Deciphering Gene Regulation from Time Series Data

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

doctor rerum naturalium (Dr. Rer. Nat)
im Fach Physik

eingereicht an der
Mathematisch-Wissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von
Dipl.-Phys. Sabrina Hempel

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Wissenschaftlichen Fakultät I:
Prof. Dr. Andreas Herrmann

Gutachter:
1. Prof. Dr. Jürgen Kurths
2. Prof. Dr. Ulrich Parlitz
3. Dr. M. Carmen Romano

eingereicht am: 27.01.2012
Tag der mündlichen Prüfung: 01.10.2012

Contents

Zusammenfassung in deutscher Sprache

Netzwerke sind allgegenwärtig in unserem Leben und können die verschiedensten Systeme repräsentieren, angefangen von sozialen Netzen über Verkehrs- und Handelsnetze bis hin zu biologischen Netzwerken, wie z.B. metabolische oder genregulatorische Netze. Die Netzwerktheorie ermöglicht den Vergleich sehr unterschiedlicher Systeme und kann somit neue Einblicke in die allgemeinen Eigenschaften und die Dynamik des untersuchten Systems liefern. Da Informationen über die Kopplungsstrukturen jedoch häufig stark limitiert sind, ist die datengetriebene Rekonstruktion der Netzwerke ein entscheidender Forschungsgegenstand. Dies erfordert die Analyse multivariater, meist zeitaufgelöster Daten, um die Kopplungsstrukturen ableiten zu können.

Meine Arbeit beschäftigt sich vorrangig mit der Rekonstruktion genregulatorischer Netze, um die Funktionalität von Organismen und ihre Reaktionen auf externe Einflussfaktoren, wie z.B. Änderungen in der Lichtintensität, der Temperatur oder der Wasser- und Nährstoffversorgung, zu verstehen. Da die Komplexität der Organismen in erster Linie durch die vielfältigen Regulationsmechanismen begründet ist, erfordert dies ein umfangreiches mechanistisches, qualitatives und quantitatives Verständnis des gesamten regulatorischen Netzwerkes.

Im Gegensatz zu typischen Situationen in der Physik handelt es sich bei den Messungen der Genexpression, die als Indikator für die Kopplungsstrukturen verwendet werden, um sehr kurze, verrauschte, oft nur grob und ungleichmässig abgetastete Zeitreihen, welche meist eine Überlagerung von verschiedenen internen und externen Einflüssen widerspiegeln. Zudem ist die Anzahl der wechselwirkenden Elemente (hunderte Gene) im Allgemeinen deutlich grösser als die Zahl der gemessenen Zeitpunkte (im Durchschnitt etwa 10 pro Gen) und es sind nur wenig Realisierungen des Experiments verfügbar.

Die Rekonstruktion des Netzwerks ist daher in der Regel ein schrittweiser Prozess, wobei die Analyse von kurzen, zeitaufgelösten Daten erste wesentliche Einblicke in mögliche Wechselwirkungskreisläufe liefern kann. In meiner Arbeit beschäftige ich mich mit diesem ersten Schritt des Rekonstruktionsprozesses und habe untersucht, ob die Zeitreihenanalyse geeignete Werkzeuge zur Ableitung generegulatorischer Netzwerke bereit hält.

In diesem Zusammenhang habe ich den Relevanz-Netzwerk-Ansatz als besonders flexible Methode der Netzwerkrekonstruktion genauer betrachtet. Meine umfangreiche Vergleichstudie mit einer Vielzahl an Ähnlichkeitsmaßen hat gezeigt, dass auf Grund der wenigen Datenpunkte Falsch-Positiv-Raten von 30% bis 50% bei der Netzwerkrekonstruktion keine Seltenheit sind. Symbol- und rangbasierte Maße haben sich dabei als besonders geeignet herausgestellt, um die limitierten Daten zu untersuchen, wobei die Letzteren deutlich robuster gegenüber dem Einfluss von Rauschen sind.

Zusätzlich habe ich verschiedene Bewertungssysteme untersucht und gezeigt, dass diese die Netzwerkrekonstruktion weiter verbessern. Insbesondere, die von mir eingeführten asymmetrischen Bewertungsschemata liefern hierbei einen wichtigen Beitrag, da die meisten Ähnlichkeitsmaße symmetrisch sind und somit die Ableitung gerichteter Netzwerke nicht ohne Weiteres

ermöglichen.

Weiterhin habe ich $IOTA$ (inner composition alignment) als ein neues asymmetrisches, permutationsbasiertes Ähnlichkeitsmaß eingeführt, welches ein effektives Werkzeug zur Rekonstruktion gerichteter Netzwerke ohne die Verwendung zusätzlicher Bewertungsschemata darstellt. In meiner Arbeit habe ich verschiedene Modifikationen des Maßes (z.B. bidirectional inner composition alignment oder partial inner composition alignment) und deren Eigenschaften untersucht.

In einer umfangreichen numerischen Studie habe ich gezeigt, dass $IOTA$ geeignet ist, um statistisch signifikante gerichtete, nichtlineare Kopplungen in verschiedenen Zeitreihen (autoregressive Prozesse, Michaelis-Menten Kinetik und chaotische Oszillatoren in verschiedenen Regimen) und Autoregulation zu identifizieren. Dabei ist deutlich geworden, dass, insbesondere bei kurzen Zeitreihen und kleinen Zeitverzögerungen im System (Regulierungszeiten), die Leistungsfähigkeit von $IOTA$ bezüglich der Netzwerkrekonstruktion deutlich besser ist als die der Rangkorrelationen.

Weiterhin erlaubt $IOTA$, ebenso wie die Korrelationsmaße, die Spezifizierung des Types der Regulation (Aktivierung oder Unterdrückung), was es zu dem einzigen Maß macht, dass die Ableitung aller für die Rekonstruktion genregulatorischer Netzwerke erforderlichen Kenndaten ermöglicht. Darüber hinaus habe ich gezeigt, dass die Netzwerkrekonstruktion mit $IOTA$ kaum von der Wahl des Schwellwertes abhängt, was von besonderem Wert bei der Anwendung auf experimentelle Daten ist.

Schließlich, habe ich den Relevanz-Netzwerk-Ansatz zusammen mit dem neuen Ähnlich-keitsmaß $IOTA$ verwendet, um ein genregulatorisches Netzwerk für die Grünalgenart *Chlamydomonas reinhardtii* unter Kohlenstoffmangel abzuleiten. Dabei standen nur sehr kurze Zeitreihen der Genexpression zur Verfügung. Das rekonstruierte Netzwerk bildet die Grundlage für weitere Experimente, um ein genaueres Bild der Funktionalität auch höherer Pflanzen zu erhalten, da es die Identifizierung von Genen ermöglicht, welche eine Schlüsselrolle bei der Regulation des Kohlenstoffkonzentrationsmechanismus in *Chlamydomonas reinhardtii* übernehmen.

Introduction

Networks are ubiquitous in nature [AB02, DZD+10] ranging from power grids [AAN04, CLM04] and representations of the climate system [DZMK09], to socio-economic [JW02], trade [SW92], and transition networks [NF08, NDSS11], to neural [DAB03, DCB07, ZZZL+07], metabolic [FFF+03], and gene regulatory networks [dJ02], just to name a few. Originating from the development of graph theory in the 18th century, the comprehensive analysis of complex networks commenced in the 20th century [WS98, BRV01, MSOI+02, WSS02, AB02, XBS02, Alb05, YP11]. Although it is a rather new approach in science, with the earliest applications in political economy and sociology [BLM+06], network theory has become a useful tool in a wide variety of scientific disciplines, including social, economic, natural and computer sciences.

During the last few decades network approaches have been widely used in this context to gain new insight into the properties and dynamics of various systems. Moreover, recent studies imply that reconstructing the network topology (*i.e.*, understanding the various patterns of interaction among the elements of the investigated system) already provides significant insight into the general dynamical behavior of the system [ZZXS10]. Hence, the data-driven reconstruction of (directed) networks is a pressing research problem with valuable applications.

This reconstruction renders the analysis of multivariate time-resolved data crucial in identifying couplings (so-called "links" in a network) and understanding the drive-response relationships, since the available knowledge about the underlying network topology and dynamics is often limited. The study of such time series data to infer couplings is very common in physics and economics (*e.g.*, to analyze stock prices, or temperature values); however, it is also becoming increasingly more prominent in chemistry and biology. In this context, data sets of interest are, for instance, time series of chemical oscillators whose drive-response relationships need to be uncovered. An other example is that of gene expression measurements, from which the topology of gene regulatory networks shall be determined. The nodes of these networks represent genes which regulate each other as well as specific functions of the organism, such as the circadian clock, cell division, or leaf growth and flowering of plants.

In this thesis, I focus on the particular problem in **understanding gene regulatory networks**, since they are essential to uncover the functionality of an organism and its response to external influences. Due to the constantly changing environmental conditions, organisms must show robustness with respect to different external stimuli (*e.g.*, changes in light intensity, temperature, or water and nutrient supply). Hence, understanding the basic genetic principles of regulation and the adaptation mechanisms is a crucial problem of current interest.

Questions of general interest include: How will various organisms react under climate change? How can crops be made more productive and robust, avoiding at the same time negative effects on other plant functions or the consumer, to ensure food supply for a fast growing world population? How can we cure diseases or at least minimize there spread? To address these and similar

questions it is crucial to understand the biology at the system level: only a systematic approach to the problem enables modeling of the complex system, identifying possible stable states and predicting future behavior under distinct external conditions.

Recent evidence from fully-sequenced genomes suggests that the complexity of an organism arises more from the elaborate regulation of gene expression than by the genome size itself. Thus, understanding the regulatory mechanisms requires not only unraveling the genetic code, but also comprehensive mechanistic, qualitative and quantitative insights into the gene regulatory networks of distinct organisms. Currently, however, such a detailed analysis is typically limited for a number of reasons:

(i) Most genomes appear as huge networks of genes which are more or less susceptible to environmental factors. Hence, parallel measurements of gene expression by advanced (so-called "high-throughput") technologies are required.

(ii) Despite the decreasing costs of high-throughput experiments, systems biology studies produce relatively short time series with an average of only 10 time points per gene, largely due to the problems with gathering a big enough sample material.

(iii) The read-outs of the system reflect the fact that the expression of one gene usually represents a superposition of various internal and external influences.

(iv) The reconstruction and modeling of the regulatory network are additionally aggravated by the fact that measurements of gene expression are noisy, and incorporate only few realizations (replicates).

(v) Time-resolved biological experiments usually involve sampling at irregular rates in order to capture processes which happen at different time scales.

Thus, the currently employed network reconstruction methods cannot directly infer the correct gene regulatory networks, due to the limited data availablity. Usually, the network inference is a stepwise process with an interplay of experimental and mathematical analyses in order to first reconstruct the network topology and then model the dynamics. In the general case, the initial network reconstructions, however, suffer from large false positive rates of approximately 30% to 50%. Hence, understanding the complex network of the entirety of gene regulatory interactions from a given system read-out necessitates the design, analysis, and testing of *new* network inference techniques (*reverse engineering* methods).

Therefore, the aim of my thesis is to compare and evaluate different tools for the network reconstruction and to develop an **improved** method for a **data-driven topological reconstruction** of complex gene regulatory networks from short gene expression time series, which yields **lower false positive rates**. The capabilities of a generalized relevance network approach are studied as a flexible *reverse engineering* method using various association measures and scoring schemes described in detail in the related chapters. Additionally, a novel association measure is introduced and discussed.

My work is structured as follows:

In Chapter 1, I briefly elucidate the basic biological background of gene regulation and explain the terms "gene expression" and "gene regulation". Furthermore, general findings on the topology of gene regulatory networks are summarized from the contemporary literature. Moreover, the problems of revealing complex gene interactions from time-resolved data of gene expression are elucidated. I discuss several approaches (clustering, topological reconstruction, and dynamical modeling) for reconstructing reverse engineered gene regulatory networks.

Next, I introduce the generalized relevance network approach and discuss how it can be used as a tool for topological reconstruction. A **systematic review** on the capabilities of that approach applied to short gene expression time series is given, using different association measures (Chapter 2) and scoring schemes (Chapter 3). Additionally, I discuss various influences on the reconstruction efficiency in Chapter 4.

In Chapter 5, I introduce a **novel permutation-based measure**, named "inner composition alignment" ($IOTA$), in order to infer directed networks from short time series without additional application of scoring schemes. The measure is compared to the association measures introduced before, and its invariance structure is investigated. In particular, the features of the **novel** measure are studied in detail with numerical simulations in Chapter 6 (for small network modules) and Chapter 7 (for gene regulatory networks).

Finally, I apply $IOTA$ within the relevance network approach to analyze empirical multivariate time-resolved gene expression data of the green algae *Chlamydomonas reinhardtii* under changing environmental conditions. The reconstructed network can serve as a basis for further experiments, since this **new measure** yields network reconstructions which are less sensitive to the particular choice of a threshold and have lower false positive rates than obtainable with the standard association measures. Candidate genes are identified which may play an important role in the adaptation of plants to changing CO_2 conditions.

1 Background

In order to outline the problem of reconstructing gene regulatory networks (GRN's), I briefly summarize the basic biological background of gene regulation. Furthermore, I present general findings on the topology of GRN's and approaches for their reconstruction, as well as the current state of gene expression data assessment.

1.1 Basic genetic principles

In most organisms, except for a few microorganisms[1], the genetic information is stored in the deoxyribonucleic acid (DNA) as an encoded robust plan – the genetic code – made up of four nucleobases: adenine (A), guanine (G), cytosine (C), and thymine (T). Each nucleobase is attached to a five-carbon sugar molecule (2'-deoxyribose) together with one to three phosphate groups, forming monomers called nucleotides, which in DNA are arranged in two long polymer strands as a double helix. The strands contain complementary bases and are held together by chemical bonds with A bonding only to T, and C bonding only to G. Furthermore, the double helix is coiled such that the genetic information is not constantly directly available, since particular segments of the DNA are not accessible for molecules to bind.

The specific sequence of the nucleobases provides the information for building and maintaining the organism, where distinct segments of the DNA are responsable for distinct functions of the organism. In that context, a gene is a segment of the DNA that contains the information on the chemical composition of a particular ribonucleic acid (RNA[2]) or protein[3].

The expression of genes, *i.e.*, the transcription[4] and translation[5] of DNA sequences, results

[1] Several plant viruses, many animal viruses and a few bacteriophages do not rely on the stable deoxyribonucleic acid (DNA), but on ribonucleic acid (RNA) as genetic material, where the nucleobase thymin (T) is replaced by uracil (U).

[2] RNA molecules play an active role in cells by catalyzing biological reactions, controlling gene expression, or sensing and communicating responses to cellular signals. The organisms use messenger RNA (mRNA) to carry the genetic information that governs the synthesis of proteins. Moreover, there are non-coding RNAs, *e.g.*, tRNAs and rRNAs. Transfer RNA (tRNA) molecules are used to deliver amino acids to the ribosome, where ribosomal RNA (rRNA) links amino acids together to form proteins.

[3] A protein is an organic compound made of amino acids arranged in a linear chain and folded into a globular form. Proteins are either used to construct the cell or take part in signaling and control of processes that are necessary for life (*e.g.*, as transcription factors, enzymes which are catalyzing metabolic reactions, or components of signal transduction pathways, to name just a few).

[4] Transcription is the process of creating an equivalent RNA copy of a sequence of DNA which acts as a master copy. For genes encoding proteins transcription results in messenger RNA (mRNA), which will be further translated into proteins.

[5] Translation is the synthesis of a protein using a mRNA molecule as template. The process incorporates mainly mRNA, tRNA and rRNA.

in biochemical material which is called the gene product. In the general case, this term refers to proteins, which may fulfill either structural, enzymatic or gene regulatory tasks. However, genes may also encode different RNAs. A gene is termed "expressed" in a particular cell and at a specified time, if the gene product it encodes is actually synthesized. Thus, the expression profile, *i.e.*, the expression level of all genes, reflects the current state of the cell. Furthermore, the analysis of the time evolution of the expression profile reveals possible regulatory pathways.

In 1958 Francis Crick elaborated the idea that genetic information flow in cells is essentially one-way, nowadays known as the central dogma of molecular biology: Information cannot be transferred back from a protein to either another protein or a nucleic acid (DNA or RNA). Therefore, the general information transfer includes only three processes:

(i) Replication[6] (information flow from DNA to DNA),

(ii) Transcription (information flow from DNA to RNA), and

(iii) Translation (information flow from RNA to protein).

With this in mind, gene expression can be regarded as the essential process to understand the adaption of organisms to changing environmental conditions. Hence, it is not surprising that the determination of interactions between genes, which govern particular system's function and behavior via gene expression, represents the grand challenge to understand the basic priciples of life [SMC07].

The process of gene expression, as sketched in Fig. 1.1, can be regulated at several stages governed by specific proteins which bind at distinct locations to the DNA and influence – direct or indirect – the transcription and translation of genes. Although the regulation is more complex in eukaryotic[7] organism than in prokaryotic[8] ones, the basic mechanisms are similar.

In particular, the control of the transcription rate, as the first step of a huge regulatory machinery, emerges in both types of organisms, governing when the transcription of a gene takes place and how much RNA is produced during the transcription process. The transcription rate is adjusted by transcription factors (master control proteins) that bind directly to DNA sequences. Moreover, it is modulated by transcription regulator proteins. In many cases, the reconstructed network will include only this transcriptional regulation, since experimental measurement techniques often do not detect the final gene product, but one of the precursors (typically mRNA).

[6] An important property of DNA is that it can replicate itself, where each strand in the double helix can serve as a template for duplicating the sequence of bases. The complex process of DNA replication is governed by specific gene products.

[7] An eukaryote is an organism whose cells contain complex structures inside the membranes. Eukaryotic cells have a nucleus (kernel) within which the genetic material is carried. In eukaryotic cells transcription is localized in the nucleus, while translation takes place outside. Thus, some RNA processing is necessary in these cells, because transcription and translation do not occur in the same place.

[8] A prokaryote have few internal structures that are distinguishable under a microscope. In particular, prokaryotic cells lack a membrane-bound nucleus (kernel). Transcription and translation occur at the same location and translation starts even before transcription is finished. Hence, RNA processing does not happen in prokaryotic cells

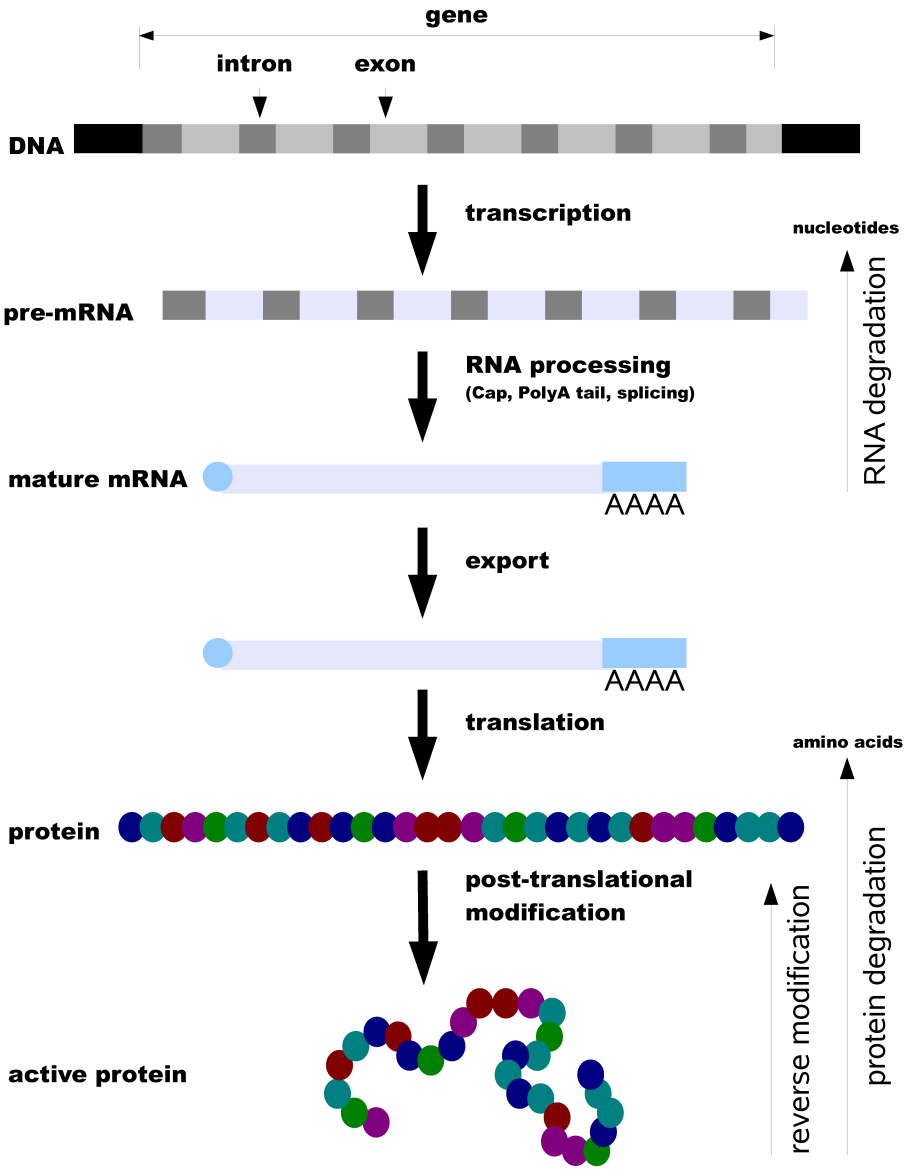

Figure 1.1: Scheme of gene expression and its main regulatory mechanism in eukaryotic cells where the DNA is composed of exons (expressed regions) and introns (intragenic regions). Several post-transciptional modifications (processing steps, *e.g.*, capping, polyadenylation or splicing) are needed to convert the primary transcript (pre-mRNA) into the mature RNA which carries the genetic information to the cytoplasm where translation takes place. In prokaryotic cells there are no introns and transcription and translation occur at the same place. Thus, RNA-preprocessing and export are not needed.

1.2 Gene regulatory networks (GRN's)

The expression of one gene can be regarded as a combinatorial action of gene products[9], whose occurrence is controlled by the expression of other genes or even the gene itself (autoregulation[10]). With this in mind, the genes can be associated with the nodes of a complex network[11] (graph), which is a collection of nodes (vertices) and links (edges) connecting pairs of nodes. In general, networks can be either undirected, meaning that there is no distinction between the two nodes associated with each link, or its links may be directed from one node to another, which is the case for regulatory networks.

Since recent studies suggest that the complexity of organisms arises basically from the regulation of the assembly of gene products via gene expression [LT03], these regulatory processes can be interpreted as the directed links in the GRN. In this context, a link from gene k to gene l means: In the particular cell / tissue there is a gene product of k that governs the expression of gene l. Thus, measured gene expression levels can be used as an indicator to determine the existing interactions.

However, the reconstruction of the GRN is difficult, since the expression of one gene usually represents a superposition of various internal and external influences. Additionally, the inference of regulatory links is aggravated by the fact that gene expression measurements usually incorporate only very few time points, often sampled at irregular rates. Hence, the elucidation of a complete network of regulatory interactions parameterized with kinetic information leading to a particular gene expression is, at present, still a challenging task. This is true even for the well-studied model organisms whose networks have been partially assembled for a few selected processes, conditions, or on the level of the entire genome [DRO$^+$02, LRR$^+$02, SOMMA02a, ZLS$^+$06].

1.2.1 What is know about the network properties?

The ever increasing throughput in experimental manipulation of gene activity in combination with the methods for quantitative assessment of transcriptome[12], proteome[13], and metabolome[14] have begun to identify the effects of individual transcription factors, binding ligands[15], and post-translational modifications on regulated genes [BS05]. Moreover, such high-throughput transcriptomics data sets can be used to identify gene regulatory modules. Nevertheless, the elaborate GRN's of most organisms are still barely understood.

Hence, in general the underlying network and its properties are not known prior to the reconstruction process. However, the available experimental and theoretical research suggests that

[9] In the context of gene regulation, one distinguishes between cis-regulatory elements (a region of DNA or RNA that regulates the expression of genes located on that same molecule) and trans-regulatory elements (diffusible factors which may modify the expression of distant genes).

[10] Autoregulation [Alo07] is an internal adaptive mechanism by which a subsystem regulates itself.

[11] Complex networks are all networks with non-trivial topological features.

[12] The term transcriptome refers to the set of all RNA molecules in a cell at a particular time.

[13] The proteome is the entire set of proteins which are expressed in a cell type or organism at a specific time point and under well-defined conditions.

[14] The metabolome represent the complete set of small-molecule (metabolites) within a biological sample which are products of chemical reactions that happen in living organisms to maintain life.

[15] Binding ligands are substances that form a complex with a biomolecule to serve a biological purpose.

GRN's, which in first approximation can be regarded as transcriptional regulatory networks, most likely are characterized with scale-free properties [Alb05].

In particular, the total degree distribution[16] was observed to follow a power law $p(k) \propto k^{-\gamma}$ with an exponent $2 < \gamma < 3$. However, more detailed studies revealed that basically only the out-degree distribution $p_{out}(k)$ shows a power law behavior ($1 < \gamma < 2$), whereas the in-degree distribution is better approximated by an exponential distribution $p_{in}(k) \propto e^{-\gamma k}$ [SSP09]. Furthermore, the distribution of the local clustering coefficients[17] exhibits approximately a power law with exponent around 1, which is not generic for scale-free networks in general, but reproduced by hierarchical network models [AVB07].

1.2.2 Revealing gene interactions – from clustering to dynamics

In order to understand the complex network of gene regulatory interactions from a given transcriptome read-out it is crucial to design, analyze, and test *reverse engineering* methods, which may, determined by the quality and quantity of the data, incorporate clustering, topological reconstruction, or dynamical modeling to gain a deeper insight into an organism's functionality.

Although, recent studies imply that reconstructing the network topology already allows to gain significant insight into the general dynamical behavior of a system, this information alone is not sufficient to quantitatively predict future gene expression levels. In addition to structural information regarding the regulatory interactions, a comprehensive understanding of the dynamical behavior of these interactions requires the specification of the type of regulation (*i.e.*, activation or inhibition) [AO03], the kinetics of interactions [RRSA02], and the specificity of the interactions with respect to the investigated tissue and/or stress condition [LBY$^+$04].

Thus, reverse engineering the entirety of GRN's of distinct organisms is a stepwise process, which includes:

(i) the application of gene clustering algorithms to identify genes with similar functionality,

(ii) topological reconstruction to infer (causal) relationships and the role of distinct genes within the large GRN, and

(iii) dynamical modeling of the regulatory subunits in order to predict future behavior.

The best reconstruction and modeling results can be achieved if complementary methods (experimental and analytical) are combined.

Clustering algorithms

In a first step clustering algorithms are frequently applied in order to identify co-regulated genes and facilitate, *e.g.*, to trace back the recent evolutionary history of an organism or to infer unknown functions of a gene and its product from the knownledge of functions of co-regulated genes. Various algorithms have been developed for this task, which can be either centroid-based,

[16] In network theory, the degree of a node is the number of connections it has to other nodes and the degree distribution is the probability distribution of these degrees over the whole network.

[17] The local clustering coefficient of a node in a network quantifies the probability of the adjacent nodes to be connected to each other as well.

connectivity-based or even distribution-based.

One of the most common algorithms is the centroid-based K-means method [Mac] which aims to partition the points of a data set into k groups by minimizing the distances from all points to randomly distributed cluster centers (centroids) to which they are assigned. The self-organized map (SOM) method [Koh82] is closely related to the previous one, however, in the beginning the cluster centers are located on a predefined grid which is deformed during the iteration process to fit the data.

Another approach are hierarchical clustering algorithms which are connectivity-based and can be divisive or agglomerative, where the latter ones are more common. Hierarchical agglomerative clustering starts from a situation where each cluster contains exactly one gene. In the standard form [Joh67], at each iteration the two closest clusters (distance of the centers) are merged into a larger one. In contrast to the previous two methods, the number of clusters is not predefined here.

A modified version of this, commonly used for gene expression analysis, is the quality threshold clustering [HKY99] where each gene in turn plays the role of a candidate gene. Iteratively genes that minimize the increase in the cluster diameter are added to the candidate gene's cluster as long as the diameter does not exceed a predefined diameter threshold (quality threshold). Then, those genes which are merged in the largest cluster are removed from the list of genes and the clustering process is repeated with the rest of the list until the largest remaining candidate cluster includes fewer than a predefined number of genes.

The most prominent distribution-based clustering method is known as expectation-maximization algorithm (or short: EM-clustering) [DNR77], where the data set is modeled with a fixed number of probability distributions (usually Gaussian distributions). The parameters of the distributions are iteratively optimized. In contrast to centroid-based or connectivity-based approaches, for each cluster membership functions are defined which allows the genes to be part of several clusters.

Each of the clustering approaches has its merits and drawbacks and require specification of different parameters, *e.g.*, the number of clusters (K-means), the number of iterations (hierarchical clustering) or the distribution of the data (EM-clustering). Thus, the obtained clusters may differ among the methods. The choice of a clustering method depends on the data and the available *a priori* knowledge.

Topological reconstruction

After potentially relevant genes for the problem under study have been identified, their interaction structure needs to be determined. Graphical modeling approaches, such as relevance, regression or Bayesian networks focus on the topological reconstruction of the underlying GRN, in order to find generic features by analyzing, *e.g.*, statistical graph properties, or community structures.

Relevance and regression networks lead to very similar represention of the system as a network, where nodes represent variables (genes) and edges represent hypothetical associations between these variables.

To assign the edges in relevance networks [BTS+00] a specific measure (usually mutual information or correlation) is chosen to estimate the dependency between all pairs of variables. Next, a particular threshold τ is set in order to account for "true" links between elements in the network. This algorithm can be further generalized by using partial and conditional association measures instead of pairwise ones, or applying a scoring scheme on the matrix of hypothetical associations.

While in relevance networks gene-gene dependencies are inferred by computing similarity scores for each pair of genes, in regression networks sets of dependencies between a target gene and all possible input genes are estimated using (linear) regression [NCARR10]. Consecutively each gene is considered as a target, those expression can be explained with different regression models (incorporating various input genes). Eventually, for each target gene the model is chosen which minimizes the error between the predicted and the actual gene expression time series. Each input gene incorporated in the chosen model is linked to the related target gene. The predicted network, however, depends on the regression model and only those functional relations can be represented which are *a priori* defined.

The Bayesian approach differs slightly in the representation of the network. A Bayesian network [Pea85] is a directed acyclic graph, where the nodes are variables[18] (gene expression values) and the edges represent conditional dependencies between the variables. Each node is associated with a probability function. These probability functions use particular sets of values for the node's parent variables[19] as input and return a probability for the variable represented by the node.

In general, in GRN reconstruction the sets of possible states are continuous and (multidimensional) density distributions have to be estimated. However, in the frequently used discret Bayesian network approach the state space is discrete or discretized and the probability distributions are given as (multidimensional) tables. The conditional dependence of subsets of variables can be further combined with an *a priori* knowledge of the underlying processes (usually a subjective expert knowledge). The graph and the conditional distributions, together defining the Bayesian network, uniquely specify a joint probability distribution.

Learning a Bayesian network can incorporate either determination of conditional (in)dependencies for a given topology, or the parameter estimation together with fitting a proper network structure. However, in both cases, infering a Bayesian network requires more information than the inference of a relevance or regression network. Additionally, it usually requires much more computational effort and is thus not feasible for the inference of large GRN's. Nonetheless it is a valuable and complementary reconstruction tool.

However, the network reconstruction often suffers from a large number of false positive links (appoximately 30% to 50%), mainly due to the very limited number of data points. Hence, cross-

[18] Variables may represent observable quantities, latent variables, unknown parameters or hypotheses

[19] Parents of a node v are all nodes from which a direct link goes to v.

validation of the predicted links with additional information from other experiments and the literature, and/or with complementary reconstruction algorithms is essential to avoid misleading interpretation of the network topology.

Dynamic modeling

While graphical modeling approaches focus on the topological reconstruction of the underlying GRN, dynamical modeling (*e.g.*, using neural network models, finite state linear models, or Boolean network models) is one step closer to biology and enables to predict future behavior.

In order to represent the continuous behavior of the gene network, neural network models [Voh01] rely on differential equations

$$\frac{d\tilde{y}^{(k)}(t)}{dt} = \frac{1}{\tau_k}\left(-\tilde{y}^{(k)}(t) + \sum_{l=1}^{N} W_{kl}\sigma[\tilde{y}^{(l)}(t-\delta\tau_{kl}) - \theta_l] + I_k(t)\right),$$

where the changes in the concentration $\tilde{y}^{(k)}$ of the product of gene k are determined by several parameters: namely a time constant τ_k, a possible external input $I_k(t)$, the shape of the activation function σ (typically a sigmoidal function), a possible offset value θ, the influence of the other genes represented by the (weighted) adjacency matrix W, and the time delay $\delta\tau_{kl}$ between the activity of gene k and gene l.

Finite state linear models [BS03] on the other hand combine continuous (protein concentration) and discret (states of the promotor region) aspects of gene regulation. In this type of models proteins attach to or detach from the promotor region if their concentration reaches a certain threshold, this process changes the state of the promotor region which corresponds to a particular level of gene activity. If a gene is active the concentration of encoded protein grows linearly depending on the activity level, otherwise the concentration decreases linearly. Finite state linear models are very simplified compared to models which rely on differential equations, however, they require less parameter estimations and thus usually less data while capturing the basic dynamical processes of gene regulation.

Boolean networks [Kau69] refer to the simplest possible dynamics on a network which facilitates easy implementation of the model and to some extend analytical investigation. Moreover, these models require less information than those which (partially) rely on the continuous data set. In Boolean networks the state of a gene is approximated with a Boolean variable (0 for expressed, 1 for not expressed) and interactions between the genes are represented by Boolean functions, where the Boolean state of a gene (a node in the network) is determinded by the states of the input genes (all nodes from which a link is directed to the considered node).

Usually the topology of the regulatory network is not known, but shall be investigated with the network model. This can be done, for example, with the *REVEAL* algorithm [LFS98]: For each gene all possible combinations of input genes are considered until a set is found which fully determines the output states, *i.e.*, the mutual information of the output given the set of inputs

equals the information content of the output alone (in terms of Shannon entropy).

However, all these dynamical models require much more experimental knowledge than the pure topological network reconstruction, while the necessary amount of data is, at present, barely available for the full set of genes even for well-studied model organisms. In particular, usually not all of the kinetic parameters are known. Additionally, in general the dynamical modelling cannot be performed for large-scale network, due to the large computational effort.

1.3 Gene expression data

Gene expression measurements under distinct, well defined external conditions enable the quantification of the level at which a particular gene is expressed within a cell, tissue or organism. However, various factors determine whether a gene is active or not, *e.g.*, the circadian rhythm, local environment, chemical signals from neighboring cells, or the phase of cell division. Furthermore, the different cell types express to some degree distinct genes. Hence, the expression analysis facilitates conclusions on the cell type, as well as on the state and environment of the cell. However, a comprehensive understanding of the functionality involves the stepwise reconstruction of the underlying GRN via reverse engineering techniques, which may operate on two different types of data sets from:

(i) *static* perturbation experiments whose read-out is a pseudo steady-state expression level, and

(ii) *time-resolved* experiments yielding time series of gene expression.

The physically represented causal structure can only be unequivocally identified by perturbing the system and observing the consequences, however, this is an expensive and time consuming task which is not feasible for all possible interactions among huge sets of genes. On the other hand, the application of association measures and network reconstruction tools on time series measurements suggests what the regulatory network may look like, and, thus, can serve as a basis to design more specific experiments. Hence, the analysis of **multivariate time-resolved** data is a crucial first step for the inference of GRN's.

Furthermore, the evaluation of existing methods for the reverse engineering process is often based on real gene expression data from high-throughput experiments. However, these data include the convoluted effects of regulons and stimulons[20], which makes it difficult to realistically assess the performance of reconstruction tools.

Moreover, not every regulatory subnetwork leads to expression of the participating genes over the measured time period and particular condition of interest. These facts lead to a lack of control when using transcriptomics time series data sets for network inference. The usage of synthetic data, in contrast to real measurements, enables a direct comparison of the performance of various reconstruction tools, since the topology and dynamic of the underlying network is known *a priori*.

[20] Regulons are genes under regulation by the same regulatory protein, whereas stimulons are genes regulated by the same external influence.

1.3.1 Experimental data

Ideally, gene expression is measured by detecting the amount of the final gene product (frequently a protein) at distinct time points, however, it is often easier to detect one of the precursors[21] (typically mRNA) to infer the gene expression level.

A common measurement technique in molecular biology is the DNA microarray, a high-throughput technology to measure expression levels of mRNA transcripts, where the relative activity of prior identified target genes is determined. For that purpose, thousands of microscopic spots of DNA oligonucleotides (short nucleic acid polymers), called features, are applied to a substrate, where each feature contains a small amount of a specific DNA sequence, called probe. The technique relies on a hybridization between two DNA strands, where a high number of complementary base pairs in a nucleotide sequence indicates tighter non-covalent bonding between the two strands.

After washing off non-specific bonding sequences, only strongly paired strands will remain hybridized, so fluorescently labelled target sequences binding to a probe sequence generate a signal which depends on the strength of the hybridization. This signal facilitates a quantification of gene expression relative to a baseline, *e.g.*, the gene expression level prior to a perturbation or a expected gene expression level.

Due to its structure a microarray enables to carry out time course experiments for large numbers of genes simultaneously. However, despite the decreasing costs of experiments relying on such high-throughput technologies, systems biology studies still produce relatively short time series [BJ04], largely due to the problems with gathering a big enough sample material and designing more complex experiments. Furthermore, correlations between gene expression time series can be induced by the specific experimental design and the essential normalization may introduce additional dependencies.

1.3.2 Synthetic data

Network generators, such as $GeneNetWeaver$ [SMF11] or $SynTReN$ [VdBVLN$^+$06a, VdBVLN$^+$06b], generate synthetic gene expression data for distinct network topologies. This facilitates to study the efficiency of different algorithms and measures for reconstructing regulatory networks from short gene expression time series.

In this thesis, I use the $SynTReN$ which creates synthetic transcriptional regulatory networks that approximate well the observed network topologies described in Section 1.2. To this end, subnetworks are extracted from well-known GRN's to provide the edges of the smaller synthetic network, as well as informations about the type of regulatory interactions, which can be either activating or inhibitory. Depending on the source network interactions can be also denoted as dual if the interaction type is not well-known or changes with changing conditions. In that case, for the generation of the simulated data the interaction type is randomly chosen as either activating or inhibitory.

Next, the generator simulates gene expression data associated with mRNA concentrations for each gene, based on Michaelis-Menten and Hill kinetics, which approximates experimental

[21] Analyzing the concentration of precursors of gene products does not capture posttranscriptional modifications which might be also important regulatory mechanisms.

expression measurements. These gene expression data sets are uniformly sampled in time. Additionally, in *SynTReN* the levels for three types of noise are user definable: (i) biological noise, corresponding to biological variability given by the stochastic variations in gene expression, (ii) experimental noise, corresponding to the technical variability, and (iii) noise on correlated inputs, which accounts for the influence of several activated genes on a regulated gene. In the general case, synthetic gene expression data with n biological and r technical replicates are generated from the particular networks.

In my investigations, I use the cluster addition strategy[22] of *SynTReN* to extract subnetworks of two well-studied model organisms: The bacterium *E. coli* [SOMMA02b] and the baking yeast *S. cerevisiae* [GBBK02].

For the simulation of the corresponding gene expression I select the same noise level for all three types of noise (the chosen values are mentioned in the related chapters) and generate $r = 6$ technical replicates. The averages of these technical replicates are used for the interaction analysis, while the biological replicates form the time series. If not stated otherwise, I use time series which consist of 10 time points, since this corresponds well to the current experimental situation. For example, in the public functional genomics data repository GEO (Gene Expression Omnibus) the largest number of time points deposited is 80, while most experiments include between 5 and 20 time points and on average approximately 10 time points are common.

[22]The cluster addition strategy, provided by *SynTReN*, has been shown to be an efficient method to extract a subnetwork that well approximates the topology of the source network [VdBVLN+06a]. In each iteration a node is randomly chosen from the source network and added to the new network together with all its neighbors.

2 Relevance networks and the question how to choose a proper measure of interaction

I focus in this thesis on the data-driven topological reconstruction of GRN's from time-resolved gene expression data using the relevance network approach. At present, time course experiments become more and more popular in molecular biology. Hence, the development and testing of methods for reverse engineering operating on time-resolved gene expression experiments is a pressing research problem. However, these gene expression measurements usually incorporate very few time points, in several cases sampled at irregular rates, which is a difficulty for reverse engineering the network.

The relevance network approach (Fig. 2.1) as a particular reverse engineering method permits to address the resulting problems by exchanging the association measure or applying an additional scoring scheme, which leads to a great flexability of the method. This generalized relevance network algorithm is based on a particular *association measure* μ and a *scoring scheme* F operating on the data matrix M (Algorithm 1):

Input:
M, matrix with q rows (genes) and n columns (time points),
μ, similarity measure,
F, scoring scheme
Output:
$q \times q$ adjacency matrix, A, of the reconstructed network G

```
1 foreach gene k, k ∈ {1,...,q} do
2 |   foreach gene l, l ∈ {1,...,q}, l ≠ k do
3 |   |   w_kl ←| μ(M_k, M_l) |
4 |   end
5 end
6 C : c_kl ← w_kl · f_kl ;
7 chose a threshold τ ;
8 a_kl ← 1 if c_kl > τ;
9 a_kl ← 0 if c_kl ≤ τ;
```

Algorithm 1: General reverse engineering method based on an association measure μ and a scoring scheme F, with $c_{kl} \in C$, $a_{kl} \in A$, and $w_{kl} \in W$. Here W is the matrix obtained by applying μ on all pairs of rows of the given data matrix M.

Let a time series profile for a gene, measured over n time points, be a sequence of expression values $y =< y_1, \ldots, y_n >$, where each y_i, $1 \leq i \leq n$, corresponds the expression at a time point t_i. Furthermore, let each of the q genes be represented by r time-resolved replicates over n time

points. In the reverse engineering process for each gene the average of these technical replicates is computed[1], resulting in a data matrix M with q rows (genes) and n columns (time points). Here, M_k, $1 \leq k \leq q$, denotes the k^{th} row of a matrix M, which corresponds to the time-resolved expression profile $y^{(k)}$ of the k^{th} gene.

Furthermore, W denotes a weighting matrix of dimension $q \times q$ that is obtained by applying an association measure μ on all pairs of rows of M, where the entries of W are $w_{kl} \geq 0$, $\forall k, l$. The entries of W indicate candidate genes likely to be regulatory related.

Moreover, the scoring scheme F is represented as a $q \times q$ matrix, and the scores obtained for a given association measure and a scoring scheme can thus be represented by a $q \times q$ matrix C calculated from the Hadamard element-wise product of W and F, such that $c_{kl} = w_{kl} \cdot f_{kl}$.

In this chapter, I investigate how an exchange of the association measure influences the capability of the algorithm to correctely infer the interrelationships between genes from time-resolved transcriptomics data. Therefore, for the moment, I will not influence the reconstruction via additional scoring. To this end, the identity scoring (ID) scheme is chosen.

2.1 The variety of association measures

I perform an extensive analysis of a set of association measures, given in Tab. 2.1, to provide a systematic review on the capabilities of the relevance network approach (Algorithm 1, Fig. 2.1), when different measures are employed to infer networks from short gene expression time series. In this chapter, only identity scoring (ID) is involved in the network reconstruction process, which corresponds to the basic relevance network approach described by Butte et al. [BK00]: Given a specific measure, the association between all pairs of genes is computed and a particular threshold τ is set in order to account for "true" links between elements (genes) in the network. The matrix F is the unit matrix ($f_{kl} = 1$) in this case.

My study includes both, common association measures and such approaches that have been borrowed from other fields. In general, given two time series $y^{(k)}$ and $y^{(l)}$ (sequences of gene expression values over n time points), an association measure is given by the mapping $\mu : \mathbf{R^n} \times \mathbf{R^n} \rightarrow I$, $I \subseteq \mathbf{R}$. In this context, the term "association measure" (μ) is a quantity that fulfills

$$\mu^{(k,l)} \leq \mu^{(k,k)}, \tag{2.1}$$

and denotes the strength of the coupling in a statistical sense. Here, $\mu(k, k)$ indicates that the two time series are identical. However, this must not be confused with an autoregulation of gene k, since pairwise measures are not capable to detect autodependencies (at least without introducing a time delay). The definition allows for the measure to be symmetric, which is, however, not commonly the case for gene regulatory interactions. The so-defined pairwise association measure detects (non)linear relationships between two variables (represented by two gene expression time series in the present study).

For reasons of comparability association measures are often normalized to the interval $[0, 1]$

[1] Alternative to the mean the meadian can be used.

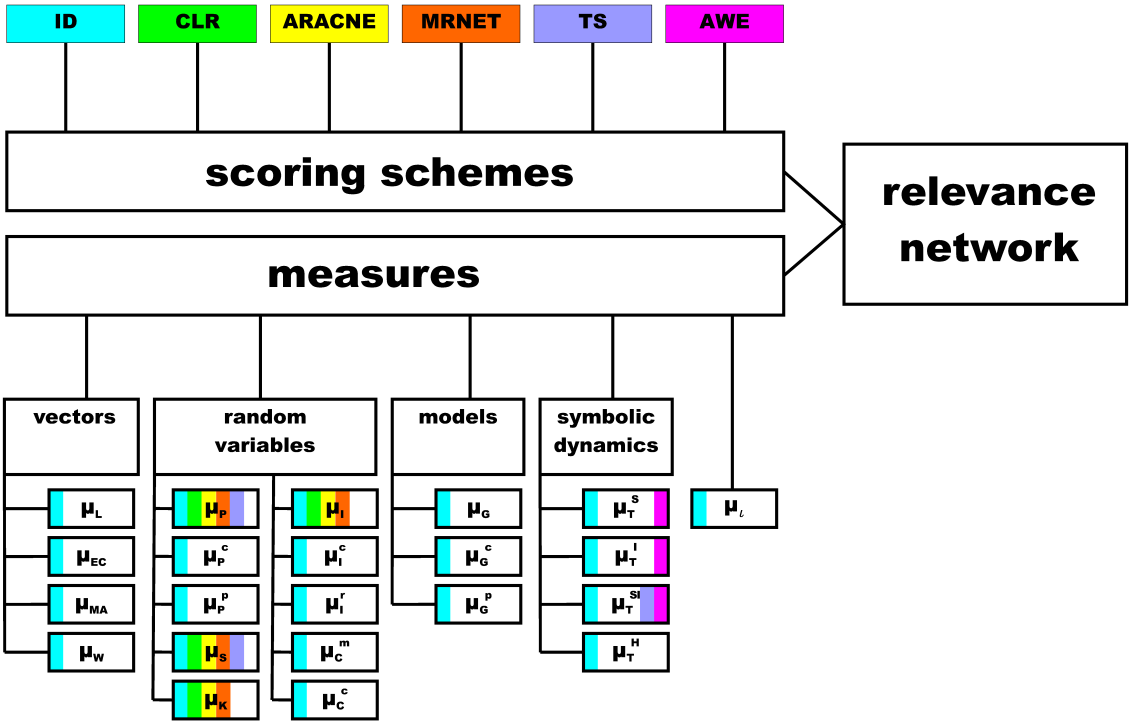

Figure 2.1: Generalized relevance network algorithm for reverse engineering GRN's. Association measures (symbols as listed in Tab. 2.1) which are common in GRN reconstruction or borrowed from other research fields are grouped based on the representation on which they operate. The different background colors indicate which combinations of scoring schemes (only short names are given here) and measures are studied. Altogether, there are 50 combinations included, since some measures can be further subdivided. Additionally, a **novel** permutation-based measure is applied together with the ID scoring scheme.

measure	symbol and reference		
	"simple" (pairwise)	conditional	partial
Euclidean distance	μ_{Ec} [WFM+98]	–	–
L^s Norm (here $s = 10$)	μ_L (in literature $s = 3$) [GR02]	–	–
Manhattan distance	μ_{Ma} [GR02]	–	–
dynamic time warping distance	μ_W [AC01]	–	–
Pearson's correlation	μ_P [ESBB98]	μ_P^c (∗) [BSS04]	μ_P^p [WZV+04]
Spearman's correlation	μ_S [DWFS98]	–	–
Kendall's correlation	μ_K [VVNV08]	–	–
mutual information	μ_I [DWFS98]	μ_I^c [LW08]	μ_I^r(*) (new)
coarse-grained information rate	μ_C^m (∗) [PKHS01]	μ_C^c (∗) [PKHS01]	–
Granger causality index	μ_G [MC07]	μ_G^c [GSK+08]	μ_G^p [GSK+08]
symbol sequence similarity	μ_T^S (∗) [MRTK07]	–	–
mutual information of symbol sequence	μ_T^I (∗) [MRTK07]	–	–
mean of symbol sequence similarity and mutual information	μ_T^{SI} (∗) [MRTK07]	–	–
conditional entropy of symbol sequence	–	μ_T^H (∗) [MRTK07]	–

Table 2.1: Overview on the association measures included in the comparison study. Those are marked by (∗) which have not been applied to gene expression data before.

and hence additionally fulfill

$$\mu^{(k,l)} \geq 0 \ \wedge \ \mu^{(k,k)} = 1. \tag{2.2}$$

However, if this is not the case, normalization is realized by dividing all $\mu^{(k,l)}$ by the largest value occurring among all pairs of genes.

If two genes are linked indirectly via a third gene, the pairwise measure can not distinguish between a direct and an indirect relationship and hence additional false positive links will be introduced in the network reconstruction. In order to reduce the number of false positive links of this type the definition of the association measure can be extended to conditional and partial measures, incorporating the possibility to exclude the influence of a third gene. However, the application of these measures for the reconstruction of GRN's is problematic:

Conditional association measures are more general, since they do not rely on specific assumptions on the probability distribution (deduced from the time series associated with a discrete random variable), instead they involve the estimation of the distribution which in turn impedes the computation of the measure from short time series. On the other hand, partial measures can indicated conditional independence reliable only for multivariate Gaussian variables. Nevertheless, to be able to discern the direction of a putative interaction and hopefully eliminate most of the spurious effects, the conditional and partial variants of the measures, if available, are considered in my study as detailed below. I term the basic pairwise measures as "simple", in comparison to their conditional and partial variants.

The evaluation of the measures is generally performed in *R* [IG09] using available packages. Additionally, several *C* routines were developed in order to improve computation speed.

The association measures, as elucidated below, can be further divided into four classes based on the representation on which they operate, namely vectors, random variables, models and symbols (Fig. 2.1).

2.1.1 Measures operating on vectors

Some of the standard measures used for determining gene regulatory interactions are based on the calculation of the distance between expression time series regarded as vectors. In what follows, $y^{(k)}$ and $y^{(l)}$ will denote the vectors $< y_1^{(k)}, \ldots, y_n^{(k)} >$ and $< y_1^{(l)}, \ldots, y_n^{(l)} >$, respectively. Various measures based on the mathematical term of the **vector norm**[2] are considered below.

L^s norm: The distance between two vectors $y^{(k)}$ and $y^{(l)}$ can be determined according to the L^s norm

$$\mu_L = \left(\sum_{i=1}^{n} |y_i^{(k)} - y_i^{(l)}|^s \right)^{1/s}. \tag{2.3}$$

In this study $s = 10$ has been chosen, which corresponds to the length n of the vectors (*i.e.*, the number of available time points). However, usually smaller values of s are considered.

[2] A vector norm denotes a function that associates a vector with a non-negative scalar. The norm is zero if the vector is a zero vector, otherwise it is positive. Furthermore, a scalar multiple to the norm is equal to the product of the norm and the absolute value of the scalar and the triangular inequality is fulfilled.

Euclidean distance: Very common is the application of the well-known Euclidian distance, which is a special case of the L^s norm, with $s = 2$. Thus, it is defined as

$$\mu_{Ec} = \sqrt{\sum_{i=1}^{n} (y_i^{(k)} - y_i^{(l)})^2}. \tag{2.4}$$

Manhattan distance: Additionally, the Manhattan distance which represents the shortest path between two points, placed on a rectangular grid, is included in this study. This distance is analogous to the L^1 norm, and is defined as

$$\mu_{Ma} = \sum_{i=1}^{n} |y_i^{(k)} - y_i^{(l)}|. \tag{2.5}$$

Dynamic time warping (DTW): Moreover, the performance of the *DTW* is investigated, which has not been applied to the problem of GRN inference before, but rather on clustering gene expression data [AC01, FC06]. The *DTW*-based measure relies on finding the optimal (*i.e.*, least cumulative) distance, mapping a given time series into a reference time series, where both sequences may vary in time and/or speed. It was originally developed for speech recognition [SC78, VZ70], but has been recently used for different data mining tasks in medicine and bioinformatics [CPB+02, AC01]. The concept is sketched in Fig. 2.2 exemplarily for two short time series with 4 time points each. In the first step of the *DTW* algorithm, local distances (*e.g.*, Euclidean or Manhattan distance) for all pairs of time points are calculated. Then, the time series are mapped into each other by linking various time points, such that each point is included at least once and the sum over the lengths of all those links is minimal (optimal alignment path). Here, the *DTW* is used as it is implemented in the *R*-package "*d*tw" [Gio09, Gio10, TGQS09], with the Euclidean as point-wise local distance, and different step patterns which indicate the local constraints of the alignment paths. The results obtained by using three different step patterns to find an optimal alignment are compared, namely:

- symmetric1

$$\mu_{W_{i,j}} = min(\mu_{W_{i,j-1}} + \mu_{Ec_{i,j}}, \mu_{W_{i-1,j-1}} + \mu_{Ec_{i,j}}, \mu_{W_{i-1,j}} + \mu_{Ec_{i,j}}), \tag{2.6}$$

  a commonly used quasi-symmetric and non-normalizable step pattern that is biased in favor of tilted steps,

- symmetric2

$$\mu_{W_{i,j}} = min(\mu_{W_{i,j-1}} + \mu_{Ec_{i,j}}, \mu_{W_{i-1,j-1}} + 2\mu_{Ec_{i,j}}, \mu_{W_{i-1,j}} + \mu_{Ec_{i,j}}), \tag{2.7}$$

  a step pattern that weights one diagonal step same as two equivalent steps along the sides and that can be normalized by dividing by the sum of the two time series lengths,

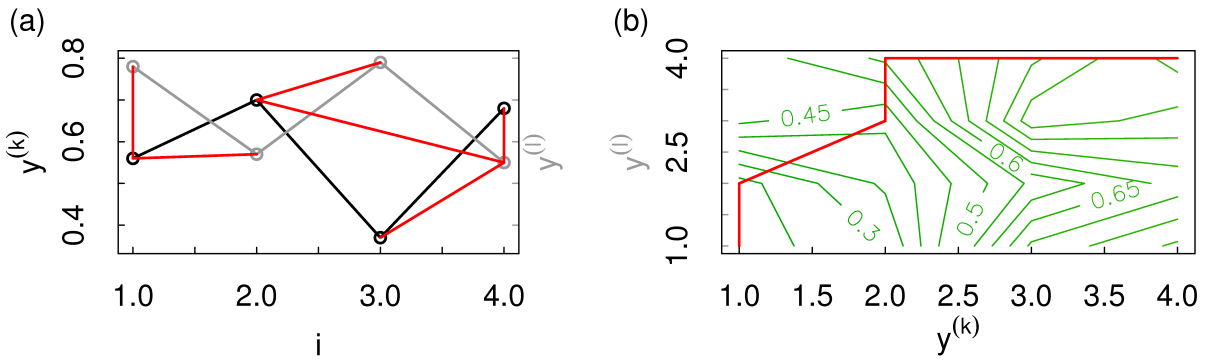

Figure 2.2: Illustration of the concept of dynamic time warping (DTW): The panel (a) shows two time series $y^{(k)}$ (black) and $y^{(l)}$ (gray), as well as a mapping (red lines) of the time points in $y^{(k)}$ into those in $y^{(l)}$. That mapping is optimal with respect to the step pattern "symmetric2". Hence, the sum of all incorporated local distances (represented by lengths of the red lines) is minimal given the constraints from the step pattern. In (b) all local distances between time points in $y^{(k)}$ and $y^{(l)}$ are shown in a contour plot, where the red path is associated with the lowest value of the cumulative distance (optimal alignment path).

- and asymmetric

$$\mu_{W_{i,j}} = min(\mu_{W_{i-1,j}} + \mu_{Ec_{i,j}}, \mu_{W_{i-1,j-1}} + \mu_{Ec_{i,j}}, \mu_{W_{i-1,j-2}} + \mu_{Ec_{i,j}}), \tag{2.8}$$

  a step pattern which is normalizable dividing by the length of the query time series. The slope is constrained between 0 and 2 and each element of the query time series is matched exactly once.

Here μ_{Ec} denotes the local (Euclidean) distance, and the measure μ_W the cumulative distance (representing the minimum sum of local distances along the alignment paths). The time index i (j) relates to the time series $y^{(k)}$ ($y^{(l)}$).

The resulting matrix of cross-distances D contains the pairwise calculated distance measures (μ_{Ec}, μ_L, μ_{Ma}, or μ_W, as defined above) and is, in all cases, normalized by the largest value occurring in the matrix:

$$D_{norm} = D/\max(D). \tag{2.9}$$

The corresponding normalized (symmetric) association measure is defined as:

$$\mu_D = 1 - D_{norm}. \tag{2.10}$$

2.1.2 Measures operating on random variables

Despite the representation of the expression time series as vectors, time series $y =< y_1, \ldots, y_n >$ can be associated with a discrete random variable Y with probability distribution $p(y)$, $y \in Y$ that is approximated by the frequency via standard binning arguments. This representation

allows to calculate several widely used association measures, such as correlation and information-theoretic measures. However, the temporal information is lost by this representation of the time series data.

Pearson correlation (PC): The use of Pearson's product moment coefficient is common to quantify the linear relationship between two random variables X and Y, corresponding to two time series $y^{(k)}$ and $y^{(l)}$. The measure is defined as:

$$\mu_P(y^{(k)}, y^{(l)}) = \frac{E[(Y^{(k)} - E[Y^{(k)}])(Y^{(l)} - E[Y^{(l)}])]}{E[(Y^{(k)} - E[Y^{(k)}])^2] \cdot E[(Y^{(l)} - E[Y^{(l)}])^2]}, \tag{2.11}$$

where E denotes expectation

$$E[Y^{(k)}] = \sum_{i=1}^{n} (y_i^{(k)} p(y_i^{(k)})). \tag{2.12}$$

If the variables are independent, the correlation coefficient is $\mu_p = 0$, but the opposite is not true, as this coefficient is sensitive mainly to linear dependencies. The simple Pearson correlation coefficient obtains values in the interval $[-1, 1]$ and is symmetric.

Conditional Pearson correlation (CPC): Substituting the expectation value (Eq. 2.12) by the conditional expectation value

$$E[Y^{(k)}|Y^{(l)}] = \sum_{i=1}^{n} (y_i^{(k)} p(y_i^{(k)}|y^{(l)})), \tag{2.13}$$

where $p(y^{(k)}|y^{(l)}),\ y^{(k)} \in Y^{(k)},\ y^{(l)} \in Y^{(l)}$ is the conditional probability distribution, yields the following definition for the CPC:

$$\mu_P(y^{(k)}, y^{(l)}|y^{(m)}) = \frac{E[((Y^{(k)} - E[Y^{(l)}|Y^{(m)}])(Y^{(l)} - E[Y^{(l)}|Y^{(m)}]))|Y^{(m)}]}{E[((Y^{(k)} - E[Y^{(k)}|Y^{(m)}])^2)|Y^{(m)}] \cdot E[((Y^{(l)} - E[Y^{(l)}|Y^{(m)}])^2)|Y^{(m)}]}. \tag{2.14}$$

Thus, the conditional correlation between the time series x and y of the corresponding genes, eliminating the influence of all other genes is defined as

$$\mu_P^c(y^{(k)}, y^{(l)}) = \min_{m \neq k, m \neq l} \mu_P(y^{(k)}, y^{(l)}|y^{(m)}). \tag{2.15}$$

Partial Pearson correlation (PPC): Analogously, one could also consider

$$\mu_P^p(y^{(k)}, y^{(l)}) = \min_{m \neq k, m \neq l} \mu_P((y^{(k)}, y^{(l)}) \cdot y^{(m)}) \tag{2.16}$$

in order to eliminate the influence of all other genes, where

$$\mu_P((y^{(k)}, y^{(l)}) \cdot y^{(m)}) = \frac{E[\mathrm{Res}(y^{(k)}(y^{(m)}))\mathrm{Res}(y^{(l)}(y^{(m)}))]}{E[\mathrm{Res}(y^{(k)}(y^{(m)}))^2] \cdot E[\mathrm{Res}(y^{(l)}(y^{(m)}))^2]} \tag{2.17}$$

$$= \frac{\mu_P(y^{(k)}, y^{(l)}) - \mu_P(y^{(k)}, y^{(m)})\mu_P(y^{(l)}, y^{(m)})}{\sqrt{(1 - (\mu_P(y^{(k)}, y^{(m)}))^2)}\sqrt{(1 - (\mu_P(y^{(l)}, y^{(m)}))^2)}}. \tag{2.18}$$

The residuals are calculated following Eq. (2.19) making a linear regression of $y^{(k)}$ (respectively $y^{(l)}$) depending on $y^{(m)}$:

$$\begin{aligned}&\mathrm{Res}(y^{(k)}(y^{(m)})) = \\ &(y^{(k)} - E[Y^{(k)}])) - \frac{E[(Y^{(k)} - E[Y^{(k)}])(Y^{(m)} - E[Y^{(m)}])]}{E[(Y^{(m)} - E[Y^{(m)}])^2]}(y^{(m)} - E[Y^{(m)}]).\end{aligned} \tag{2.19}$$

Hence, the PPC is only capable to eliminate the linear influences of the other genes.

Spearman's rank correlation (SC): Rank correlations, such as the one defined by Spearman, can be used as a more general measure of interdependence, not restricted to a linear relationship. Spearman's definition of correlation is based on the rank distribution of the expression values:

$$\mu_S(y^{(k)}, y^{(l)}) = \frac{E[(\mathrm{rank}(y^{(k)}) - E[\mathrm{rank}(y^{(k)})])(\mathrm{rank}(y^{(l)}) - E[\mathrm{rank}(y^{(l)})])]}{E[(\mathrm{rank}(y^{(k)}) - E[\mathrm{rank}(y^{(k)})])^2] \cdot E[(\mathrm{rank}(y^{(l)}) - E[\mathrm{rank}(y^{(l)})])^2]} \tag{2.20}$$

and describes how well the relation between two variables can be explained by monotonic functions.

Kendall's rank correlation (KC): Another rank correlation was introduced by Kendall and measures the similarity of the ordering of the data. It is defined as

$$\mu_K(y^{(k)}, y^{(l)}) = \frac{2(n_c - n_d)}{n(n-1)}, \tag{2.21}$$

with n_c being the number of concordant pairs, and n_d the number of discordant pairs of the rank sets.
Even though rank correlations measure a different type of relationship than the product moment correlation coefficient, they are also defined in the interval $[-1, 1]$, and are symmetric. It is common to regard the rank correlation coefficients (especially Spearman's rank correlation) as alternatives to Pearson's coefficient, since they could either reduce the amount of calculation or make the coefficient less sensitive to non-normality of distributions. Nevertheless, they quantify different types of association.

Unlike most of the association measures discussed here, correlations do not only provide an information about whether two genes are interacting, but also whether it is an activating or inhibitory relationship. As the latter information is outside of the interest of the relevance

network approach, only the absolute value (respectively the square) of the correlation coefficient is taken into account for the comparison study.

Mutual information (MI): Information-theoretic measures, same as the correlations, are defined using random variables as relevant representation for expression time series. In that context, one of the most commonly used measures for inferring interdependencies between two subunits of a system is the mutual information [SBA07]. Intuitively, MI measures the information content that two random variables $Y^{(k)}$ and $Y^{(l)}$ share. The simple mutual information can then be expressed in terms of the marginal entropies $H(Y^{(k)})$ and $H(Y^{(l)})$, and the joint entropy $H(Y^{(k)}, Y^{(l)})$ using the definition of the Shannon entropy

$$H(Y) = \sum_{i=1}^{n} p(y_i)\log(p(y_i)), \tag{2.22}$$

which quantifies the uncertainty associated with a random variable. Hence, the simple MI is a symmetric measure that is defined as:

$$\mu_I(y^{(k)}, y^{(l)}) = H(Y^{(k)}) + H(Y^{(l)}) - H(Y^{(k)}, Y^{(l)}). \tag{2.23}$$

It includes also non-linear interrelations, but as the other simple measures, the simple MI cannot be used to distinguish between direct and indirect relations.

Conditional mutual information (CMI): If the marginal and joint entropies are replaced by the conditional analogs, namely $H(Y^{(k)}|Y^{(m)})$, $H(Y^{(l)}|Y^{(m)})$ and $H(Y^{(k)}, Y^{(l)}|Y^{(m)})$, the influence of a third variable Z can be eliminated:

$$\mu_I(y^{(k)}, y^{(l)}|y^{(m)}) = H(Y^{(k)}|Y^{(m)}) + H(Y^{(l)}|Y^{(m)}) - H(Y^{(k)}, Y^{(l)}|Y^{(m)}). \tag{2.24}$$

The minimal information shared by the time series $y^{(k)}$ and $y^{(l)}$ of two genes conditioned on each $y^{(m)}$, $m = 1, \ldots, q$ has to be calculated. Hence, the conditional – sometimes also referred to as partial [FP07] – mutual information can be written as:

$$\mu_I^c(y^{(k)}, y^{(l)}) = \min_{m \neq k, m \neq l} \mu_I(y^{(k)}, y^{(l)}|y^{(m)}), \tag{2.25}$$

The degree of interaction is indicated by the values of μ_I or μ_I^c normalized by the largest value occurring among all pairs of genes.

Mutual and conditional coarse-grained information rate ($MCIR$ and $CCIR$): Further measures based on information-theoretic aspects are the coarse-grained measures. Here, instead of approximating the exact entropies of time series, relative measures of "information creation" are applied to study the interrelationship of two (sub)systems. For this purpose, the calculation of coarse-grained entropy rates [PKHS01] is used to replace the approximation of the

Kolmogorov-Sinai entropy[3]: First, a time lag γ_{max} is determined such that

$$\mu_I(y_i, y_{i+\gamma'}) \approx 0, \ \forall \ \gamma' \geq \gamma_{max}, \tag{2.26}$$

among all analyzed data sets. Then, the coarse-grained information rate (CIR) is given by the norm of the mutual information

$$\mu_C(y) = ||\mu_I(y_i, y_{i+\gamma})|| = \frac{\Delta\gamma}{\gamma_{max} - \gamma_{min} + \Delta\gamma} \sum_{\gamma=\gamma_{min}}^{\gamma_{max}} \mu_I(y_i, y_{i+\gamma}). \tag{2.27}$$

Usually, the parameter γ_{min} and $\Delta\gamma$ (difference between consecutive time lags) can be set to one, and thus the CIR becomes

$$\mu_C(y) = \frac{1}{\gamma_{max}} \sum_{\gamma=1}^{\gamma_{max}} \mu_I(y_i, y_{i+\gamma}). \tag{2.28}$$

Hence, the mutual coarse-grained information rate is defined as

$$\mu_C^m(y^{(k)}, y^{(l)}) = \frac{1}{2\gamma_{max}} \sum_{\gamma=-\gamma_{max}}^{\gamma_{max}, \gamma\neq 0} \mu_I(y_i^{(k)}, y_{i+\gamma}^{(l)}), \tag{2.29}$$

whereas the conditional coarse-grained information rate is

$$\mu_C^c(y^{(k)}|y^{(l)}) = \mu_{C0}(y^{(k)}|y^{(l)}) - \mu_C(y^{(k)}), \tag{2.30}$$

with

$$\mu_{C0}(y^{(k)}|y^{(l)}) = \frac{1}{\gamma_{max}} \sum_{\gamma=1}^{\gamma_{max}} \mu_I(y_i^{(k)}, y_{i+\gamma}^{(k)}|y^{(l)}). \tag{2.31}$$

Eventually, a normalization by the largest value occurring among all pairs of genes is performed, and the normalized coarse-grained information rates are used to indicate the degree of interaction.

A novel association measure – residual mutual information (RMI): Estimating entropies from short time series is imprecise, hence the estimation of the mutual information and, in particular, its conditional counterpart, suffers the same disadvantage. On the other hand, the simple mutual information is not able to distinguish between direct and indirect links. Therefore, in order to overcome the encountered problem, a **novel** partial measure is proposed – the **residual mutual information** defined as

$$\mu_I^r(y^{(k)}, y^{(l)}) = \min_{m\neq k, m\neq l} \mu_I((y^{(k)}, y^{(l)}) \cdot y^{(m)}), \tag{2.32}$$

[3]The Kolmogorov-Sinai entropy is the metric entropy of a dynamical system and denotes the information content that is needed to predict the position of a trajectory in the phase space (divided into D-dimensional hypercubes) at a certain time.

where

$$\begin{aligned}&\mu_I((y^{(k)}, y^{(l)}) \cdot y^{(m)}) = \\ &H(\mathrm{Res}(y^{(k)}(y^{(m)}))) + H(\mathrm{Res}(y^{(k)}(y^{(m)}))) - H(\mathrm{Res}(y^{(k)}(y^{(m)})), \mathrm{Res}(y^{(l)}(y^{(m)}))),\end{aligned} \tag{2.33}$$

analogously to the idea of partial correlation (the residuals are calculated in the same way as for the partial correlation in Eq. (2.19)). The degree of interaction in the complex network is then indicated by the values of μ_I^r, normalized by the largest value occurring among all pairs of genes. Applied to short data sets, the RMI is expected to perform much better in discriminating indirect links than the CMI, as the estimation of additional conditional probabilities is not needed. Hence, the measure is more robust to effects of small sample size. Furthermore, the performance of RMI is expected to range between those of the simple and the conditional mutual information for long time series, since in contrast to the CMI, only the linear influence of the variable $Y^{(m)}$ on $Y^{(k)}$ and $Y^{(l)}$ is eliminated.

2.1.3 Model-based measures

Additionally, the similarity of time series can be investigate based on model assumptions. Therefor, a particular model is chosen and the parameters are fitted to the data. Eventually, the similarity of the time series is evaluated by comparing the estimated parameters. One of the most popular model-based approaches is Granger causality.

Granger causality index (GC): Although already devoloped in the 1960's, the Granger causality is a rather new approach to infer GRN's. Given the time series $y^{(k)}$ and $y^{(l)}$, two linear autoregressive (AR) models are estimated, both including the past of $y^{(k)}$, and additionally, one of them including the past of $y^{(l)}$. The optimal order s of the AR model, which denotes the number of past time points which have to be included, is in this study determined by the function "*VARselect*" from the *R*-package "*vars*" [Pfa08b, Pfa08a] based on the Akaike information criterion[4] (AIC) [Aka03].

With properly selected AR models, the part of the variance in the data which is explained by one model in comparison to the other one, provides an information on the (causal) relationship. This comparison can be formulated in terms of the **Granger Causality index**, denoted by μ_G for the simple linear measure, as defined in [GSK+08, DCB06] via the covariance σ,

$$\mu_G(y^{(l)} \to y^{(k)}) = \log \frac{\sigma(u_{1j}, u_{1j})}{\sigma(u_{2j}, u_{2j})}. \tag{2.34}$$

[4]The AIC is a measure of the quality of a fit of an estimated statistical model, deduced as a tool for model selection. In the general case, it is defined as $AIC = 2u - 2\log(L)$ where u is the number of parameters in the statistical model, and L is the maximized value of the likelihood function for the estimated model.

It can be inferred from the AR models:

$$y_j^{(k)} = \sum_{i=1}^{s} a_{11i} y_{j-i}^{(k)} + u_{1j} \tag{2.35}$$

$$y_j^{(k)} = \sum_{i=1}^{s} a_{21i} y_{j-i}^{(k)} + \sum_{i=1}^{s} a_{22i} y_{j-i}^{(l)} + u_{2j}, \tag{2.36}$$

where a_{11i}, a_{21i} and a_{22i} are the parameters of the models and u_{1j} and u_{2j} represent white noise.

Conditional and partial Granger causality (CGC and PGC): Same as for the previous measures, the conditional and partial (linear) Granger causality measures as defined in [GSK+08, DCB06], are used in order to identify existing indirect relationships. Hence, the related AR models are formulated as

$$y_j^{(k)} = \sum_{i=1}^{s} a_{11i} y_{j-i}^{(k)} + \sum_{i=1}^{s} a_{12i} y_{j-i}^{(m)} + u_{1j}, \tag{2.37}$$

$$y_j^{(k)} = \sum_{i=1}^{s} a_{21i} y_{j-i}^{(k)} + \sum_{i=1}^{s} a_{22i} y_{j-i}^{(l)} + \sum_{i=1}^{s} a_{23i} y_{j-i}^{(m)} + u_{2j}, \tag{2.38}$$

for the conditional, and, in addition,

$$y_j^{(m)} = \sum_{i=1}^{s} a_{31i} y_{j-i}^{(m)} + \sum_{i=1}^{s} a_{32i} y_{j-i}^{(k)} + u_{3j}, \tag{2.39}$$

$$y_j^{(m)} = \sum_{i=1}^{s} a_{41i} y_{j-i}^{(k)} + \sum_{i=1}^{s} a_{42i} y_{j-i}^{(l)} + \sum_{i=1}^{s} a_{43i} y_{j-i}^{(m)} + u_{4j}, \tag{2.40}$$

for the partial Granger causality, with a_{11i}, a_{12i}, a_{21i}, a_{22i}, a_{23i}, a_{31i}, a_{32i}, a_{41i}, a_{42i} and a_{43i} being the parameters of the models and u_{1j}, u_{2j}, u_{3j} and u_{4j} representing noise terms.

Using the Eqs. (2.37) and (2.38), the **conditional Granger causality index** is then defined as:

$$\mu_G^c(y^{(l)} \to y^{(k)}) = \min_{m \neq k, m \neq l} \mu_G(y^{(l)} \to y^{(k)} | y^{(m)}), \tag{2.41}$$

where

$$\mu_G(y^{(l)} \to y^{(k)} | y^{(m)}) = \log \frac{|\sigma(u_{1j}, u_{1j})|}{|\sigma(u_{2j}, u_{2j})|} \tag{2.42}$$

Moreover, using the Eqs. (2.37) to (2.40), the **partial Granger causality index** is defined as

$$\mu_G^p(y^{(l)} \to y^{(k)}) = \min_{m \neq k, m \neq l} \mu_G((y^{(l)} \to y^{(k)}) \cdot y^{(m)}), \tag{2.43}$$

where

$$\mu_G((y^{(l)} \to y^{(k)}) \cdot y^{(m)}) = \log \frac{\sigma(u_{1j}, u_{1j}) - \sigma(u_{1j}, u_{3j})\sigma(u_{3j}, u_{3j})^{-1}\sigma(u_{3j}, u_{1j})}{\sigma(u_{2j}, u_{2j}) - \sigma(u_{2j}, u_{4j})\sigma(u_{4j}, u_{4j})^{-1}\sigma(u_{4j}, u_{2j})}. \tag{2.44}$$

Finally, the degree of interaction is indicated by the Granger causality index ("simple", conditional or partial) normalized by the largest value occurring among all pairs of genes.

2.1.4 Measures operating on symbolic dynamics

Despite the promising applications of interaction measures based on symbolic dynamics in various fields, they have not yet been employed for reverse engineering GRN's. For instance, in standard nonlinear time series analysis, the usage of symbolic dynamics to uncover patterns of interactions, especially from short data sets [WSR$^+$09, PBL$^+$11], has proven as a valuable tool. Therefore in what follows, the potential of symbolic dynamics for the problem at hand is explored by using the principle of order patterns. As a basis I use the principle described in [MRTK07] to transform the time series into symbol sequences, however, I modify the procedure to obtain longer sequences. In general, an order pattern Π of dimension δ is defined by the discrete order sequence of the time series y and has the length δ. Hence, the time series can be symbolized using order patterns following:

$$(y_i, y_{i-j_1}, ..., y_{i-j_\delta-1}) \rightarrow \Pi_i, \tag{2.45}$$

where j is the time lag. In terms of GRN reconstruction, a specific number of time points is chosen and those points are ranked according to the expression value in order to obtain the order pattern. Then, each possible ordering corresponds to a predefined symbol.

This concept is illustrated in Fig. 2.3 for a time series composed of $n = 4$ time points. Since time-resolved gene expression time series are very short, in the following analysis all possible combinations of the chosen number of time points are considered. For the time series of length $n = 4$ and an order pattern of dimension $\delta = 3$, symbols (order patterns Π_i) are defined for the following groups of time points: $(1, 2, 3)$, $(1, 2, 4)$, $(1, 3, 4)$ and $(2, 3, 4)$, shown in the left panels of Fig. 2.3. Next, a symbol sequence

$$S^{(k)} = (\Pi^{(k)}_{j_1}, ..., \Pi^{(k)}_{j_\eta}) \tag{2.46}$$

is defined where $\Pi^{(k)}_j$ denotes the order pattern obtained for the time series of gene k from the j-th group of time points and

$$\eta = \frac{n!}{\delta!(n-\delta)!} \tag{2.47}$$

is the length of the symbol sequence.

In this work, the dimension δ is, where not otherwise stated, chosen such that the length η becomes maximal, given a time series of length n (*i.e.*, $\delta = 5$ for $n = 10$).

Symbol sequence similarity ($SySim$): Using the above described approach, the interdependency of two genes is inferred as follows: Given a certain number δ of time points a vector P is defined which contains all possible permutations of the ranking, and a symbol (order pattern Π_j) is assigned to each of these permutations. Next, a vector $\bar{P}$ is defined, using the same symbols as for P, but assigned to the reversed ranking. Then, the pattern overlap of two symbol sequences

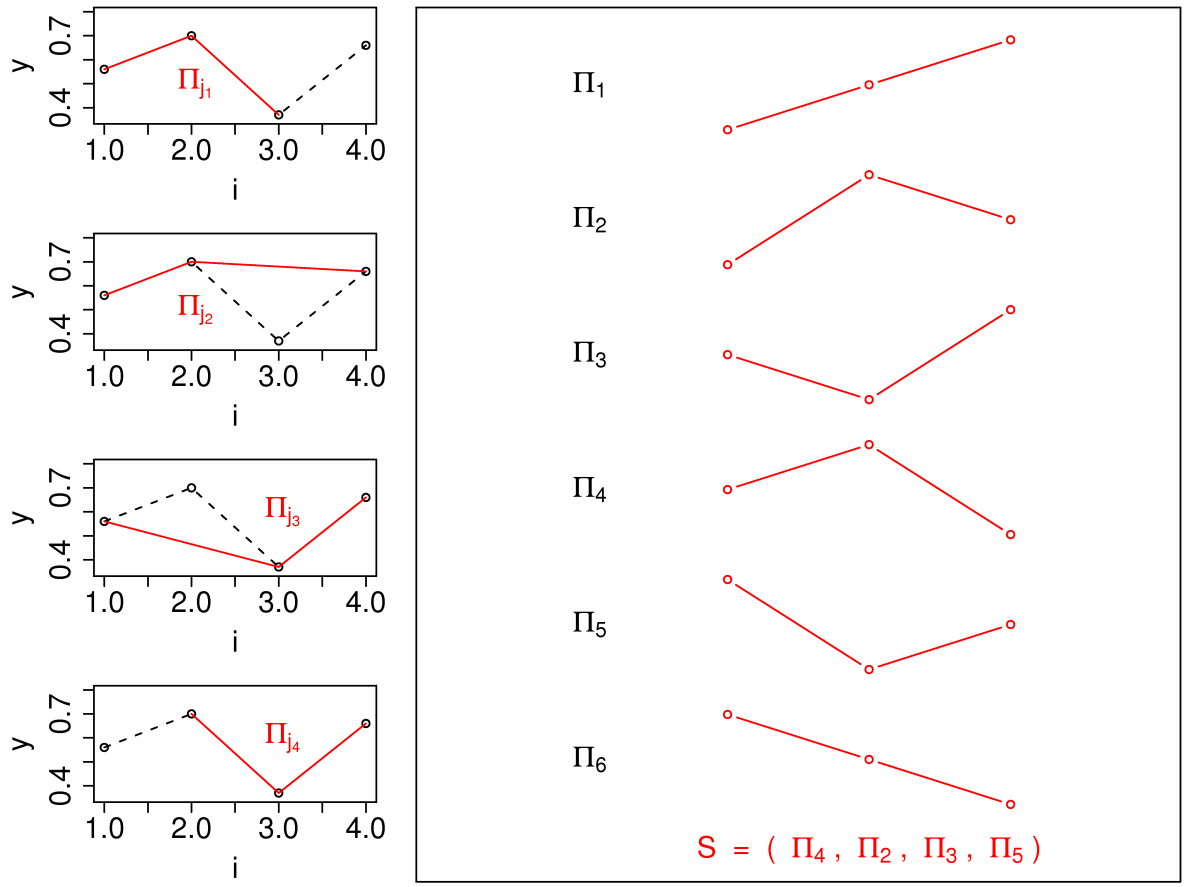

Figure 2.3: The left panels show a time series (black) composed of $n = 4$ time points and particular groups of 3 time points each which are forming order pattern of dimension $\delta = 3$ (red). An overview over the possible order pattern of that dimension is provided in the right panel together with the resulting symbol sequence S for the mentioned time series.

$S^{(k)}$ and $S^{(l)}$ is counted to evaluate the **symbol sequence similarity**, p_1, assuming both time series are interrelated (Eq. (2.48)), and p_2, if anti-interrelation (Eq. (2.49)) is assumed:

$$p_1 = \sum_{j=1}^{\delta!} \frac{\sum_{t=1}^{n_S} (S_t^{(k)} = P_j \ \wedge \ S_t^{(l)} = P_j)}{\eta}, \tag{2.48}$$

$$p_2 = \sum_{j=1}^{\delta!} \frac{\sum_{t=1}^{n_S} (S_t^{(k)} = P_j \ \wedge \ S_t^{(l)} = \bar{P}_j)}{\eta}. \tag{2.49}$$

Finally, in this work I chose the maximal value of the two frequencies p_1 and p_2

$$\mu_T^S = max(p_1, p_2) \tag{2.50}$$

to define the symbol sequence similarity.

Mutual information of symbol vectors ($SyMI$): Furthermore, the **mutual information of the symbol vectors** of maximal length

$$\mu_T^I = \mu_I(S^{(k)}, S^{(l)}) \tag{2.51}$$

is calculated to determine the interrelation of two genes.

Further measures: Based on the symbol vectors of maximal length the joint value of **symbol sequence similarity and the mutual information** of the symbol vectors ($SymSimMI$) is computed as

$$\mu_T^{SI} = \frac{1}{2} \left(\frac{\mu_T^S}{\max(\mu_T^S)} + \frac{\mu_T^I}{\max(\mu_T^I)} \right). \tag{2.52}$$

Furthermore, the study is extended to include symbolic dynamics based on a slope comparison (order patterns for pairs of time points), considering

- the **symbol sequence similarity for pairs** ($\mu_T^S \ pairs$) as a similarity measure
- and the **conditional entropies for pairs** ($\mu_T^H \ pairs$) as a distance measure, with $\mu_T^H = H(S^{(k)}, S^{(l)})/H(S^{(k)})$.

2.2 Performance of various association measures in terms of receiver operating characteristics curves

In order to provide a basis for the selection of a reverse engineering method for the network reconstruction which is suitable for given data, I compare the performance of the association measures within the classes defined above. In particular, to evaluate the reconstruction efficiency of the basic relevance network approach when distinct association measures are applied, I determine and discuss the receiver operating characteristics (ROC) curves, which are presented in the following section. These curves illustrate the change of the relative trade-offs between

benefits (true positive rate (tpr) — correctly inferred links) and drawbacks (false positive rate (fpr) — incorrectly inferred links), while continuously tuning the threshold that is used to identify a link [Faw06]. Hence, they facilitate to evaluate to which extent each of the association measures accurately reconstructs the underlying network of regulatory interactions.

I use synthetic gene expression time series[5] (10 time points simulated without noise) of 100 genes of *E. coli* for the comparison study.

2.2.1 Measures operating on vectors

Figure 2.4 (a) illustrates the efficiency of the reconstruction of links based on classical distance measures, and the dynamic time warping. In general, none of these measures is able to avoid false positives on a larger scale without loosing most of the true interactions. On the other hand, the ROC curves are rather flat for high fpr's, which implies that these measures could be useful initially to determine connections which are not present in the network. All of the curves shown in Fig. 2.4 (a) are smooth, *i.e.*, the prediction of links is not very sensitive to the explicit choice of the threshold. This insensitivity to the threshold in turn renders the distances more precise when working with experimental data.

Furthermore, it turned out that from the investigated distance measures the L^s norm (with $s = 10$, equating the length of the time series) performs best in reconstructing the network. These results outperform the Euclidean (L^2 norm) and the Manhattan (L^1 norm) distance, which can be explained by the fact that the L^s stronger weights large distances. Additionally, the dynamic time warping fails for the investigated data, which is most likely to be a result of the coarse sampling and the complexity of the network.

2.2.2 Measures operating on random variables

Next, the ID scoring scheme is evaluated using several measures which, as an argument, use time series represented via random variables. This class of measures can be further subdivided into correlation, which will be discussed first in the following section, and information-theoretic measures.

In the case of the linear Pearson correlation (PC) coefficient, as shown in Fig. 2.4 (b), almost identical results are obtained from the simple and the conditional (CPC) measure, although the CPC should eliminate indirect interactions. However, this does not mean that there are no indirect links wrongly deduced by the linear PC. The problem is caused by the estimation of the conditional probabilities, which is barely reliable from very short time series (approx. 10 time points). Even if a basic significance test is included — *e.g.*, the data is reshuffled 100 times, then the measures for the randomized series are calculated, and the obtained results are compared to those received from the original time series — the results do not differ significantly (Fig. 2.4 (d)). The partial Pearson correlation, on the other hand, shows better results for low fpr's, but looses its accuracy when high tpr's are accessed. Additionally, the results obtained from the PPC are less significant (in terms of the reshuffled time series). Removing links without significant values of the correlation yields an almost random prediction if the partial Pearson correlation is used.

[5]The data is generated as described in Section 1.3.

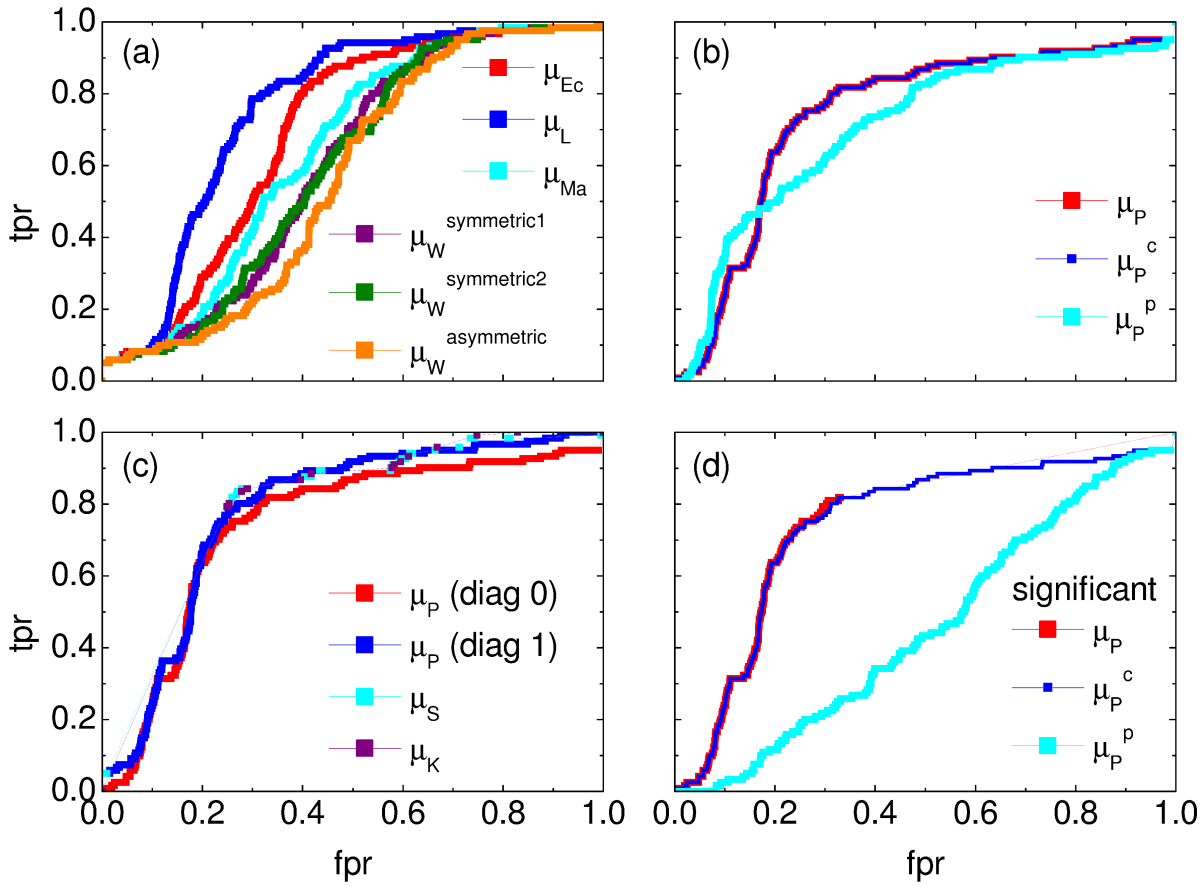

Figure 2.4: (a) ROC curves, in which the false positive rate (fpr) vs. the true positive rate (tpr) is plotted, for the network reconstruction using the identity (ID) scoring scheme and various measures operating on vectors. The results are obtained from the Euclidean distance (μ_{EC}), the L^s norm (μ_L) and the Manhattan distance (μ_{MA}), as well as from the dynamic time warping (μ_W) with the step pattern symmetric1, symmetric2 and asymmetric. (b) ROC curves obtained for the ID scoring scheme using the simple, conditional and partial Pearson correlation (μ_P, μ_P^c, μ_P^p), where the diagonal of the cross-correlation matrix is set to 0. (c) ROC curves using the ID scoring scheme and different correlation coefficients, such as the simple Pearson correlation coefficient, where the diagonal of cross-correlation matrix is once 0 (μ_P ($diag0$)), and in another case 1 (μ_P ($diag1$)). Furthermore, the ROC curves using the Spearman (μ_S ($diag1$)) and the Kendall (μ_K ($diag1$)) correlation coefficient, where the diagonal is 1 in both cases, are shown. (d) The corresponding ROC curves for Pearson's correlation, if a significance test (by reshuffling of the time series) is applied.

Since autoregulation cannot be inferred from the analysis of the similarity of short expression series with correlation measures, the diagonal of the correlation matrix was set to zero in the above computations (in general the diagonal is one since the similarity of identical time series is one per definition).

Comparing the reconstruction efficiency of the linear PC with that of the rank correlations (all diagonals equal 1), it shows up that the ROC curve in Fig. 2.4 (c) is smoother in case of Pearson correlation than the curves obtained for the rank correlations. Hence Pearson's correlation measure is less sensitive to the choice of the threshold, whereas the rank correlations can achieve a slightly better overall performance.

In the next paragraph, the efficiency of the ID scoring scheme is investigated considering information-theoretic measures. In general, the resulting reconstruction strongly depends on the method chosen for the estimation of entropies. Here the *R*-package "*infotheo*" (in particular the Miller-Madow asymptotic bias corrected empirical estimator) is used, because for short time series, as those under study, it estimates the entropy better than the *R*-package "*entropy*". Besides the basic pairwise mutual information (MI), the conditional mutual information (CMI) and the residual mutual information (RMI) are considered in order to reduce the number of false positive links. However, all these measures result in ROC curves which are more or less discontinuous. This is a finite size effect, as the time series are very short, and thus the estimation of the MI (entropies) becomes problematic.

A quite different behavior of the ROC curves is observed in specific regions of the ROC space, as shown in Fig. 2.5 (a). The simple mutual information results in a flat and comparatively smooth ROC curve for high fpr's. This means that the measure allows to remove about 60% of the false positives, by loosing approximately 10% of the true links. An even better performance in the same ROC space region can be achieved using the RMI which has been proposed here as a partial mutual information measure to distinguish indirect from direct (linear) interrelationships between triplets of genes. In contrast to this, the CMI results in a more discontinuous curve for high fpr: there, the ratio of tpr and fpr is nearly the same as observed for a random prediction. In principle the CMI is more strict in removing indirect links as it also covers nonlinear interactions. However, the conditional probabilities cannot be estimated sufficiently well from the short time series. Hence the CMI fails for the investigated short data sets in the region of high fpr's.

Additionally, when looking at the region of low fpr the ROC curve of the simple MI becomes more discontinuous than for high fpr. The tpr decreases significantly for slightly reduced threshold values, in the region around 30% and 15% of the false positives. This is manifested as jumps in the curve, due to which this measure is rather sensitive to the choice of the threshold, if low fpr's should be achieved. In contrast to this, the RMI results in a smoother curve for low fpr's than the simple MI, indicating that the measure is less sensitive to the choice of threshold, although the curve exhibits smaller jumps as well. In the region of fpr < 0.1 the performance of the RMI decreases compared to the simple measure. The CMI on the other hand, achieves only very low fpr's, which leads to low tpr (up to about 5% of true positive links). Tuning the threshold to allow for slightly higher values of the fpr the ROC curve of the CMI immediately jumps to 50% of false positives. Hence, the region between about 3% and 50% of false positive

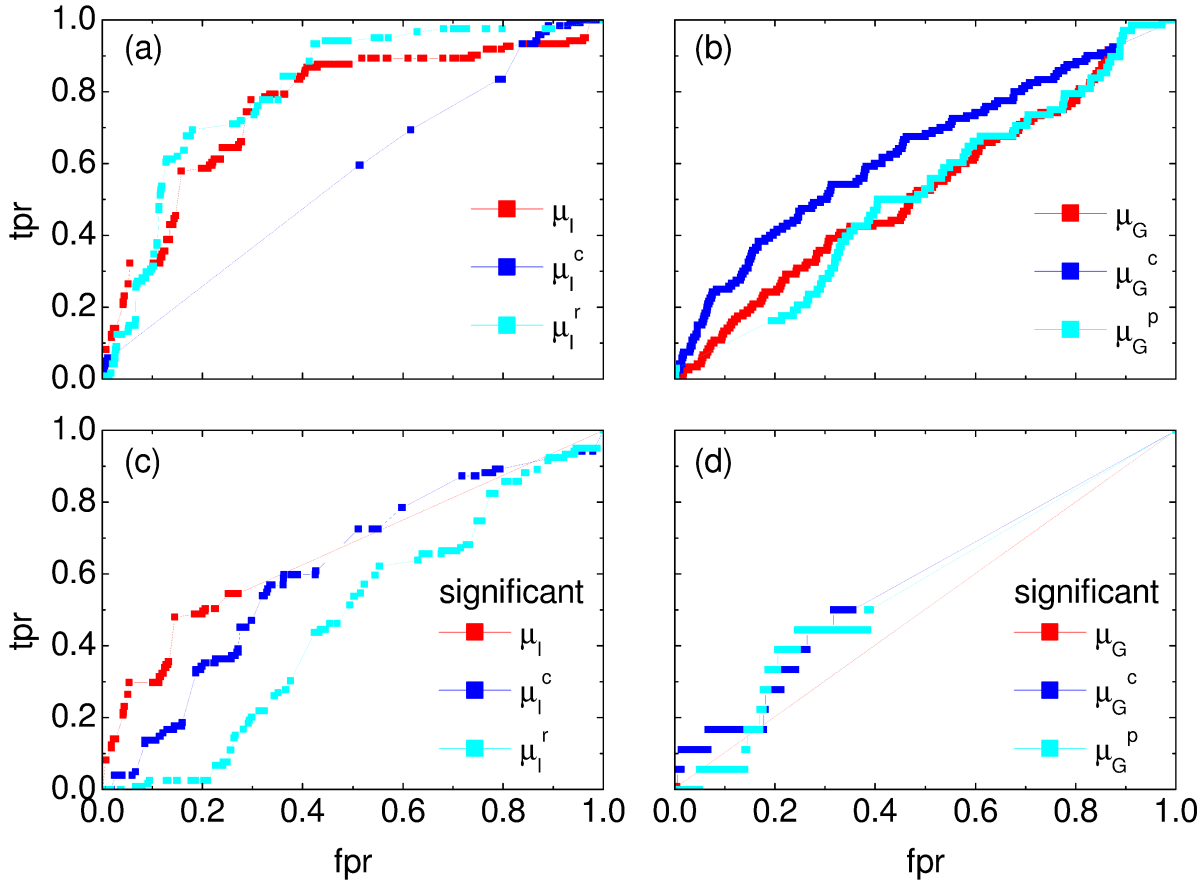

Figure 2.5: (a) Evaluation of the ID scoring scheme using information-theoretic measures: simple, conditional and residual mutual information (μ_I, μ_I^c and μ_I^r) (b) ROC curves, obtained for the simple, conditional and partial Granger causality index (μ_G, μ_G^c, μ_G^p) using the identity scoring scheme are shown. (c) ROC curves for the mutual information measures with application of a significance test by reshuffling. (d) ROC curves for the Granger causality measures with significance test.

links is not achievable using the conditional measure here.

Furthermore, a basic significance test by reshuffling the time series 100 times, calculating the measure for the randomized series, and comparing the results to those obtained for the original time series is implemented, as it was used for correlations before. The associated ROC curves for the mutual information measures after the significance test has been applied are shown in Fig. 2.5 (c). Regarding this significance information, the reconstruction efficency of the simple and the residual mutual information decreases. This is caused by the fact that the inferred interaction is, for most of the gene pairs, not significant in the specified sense. In contrast to this, if the significance test is considered, the quality of the prediction obtained from the CMI increases slightly, however its overall performance is still deficient.

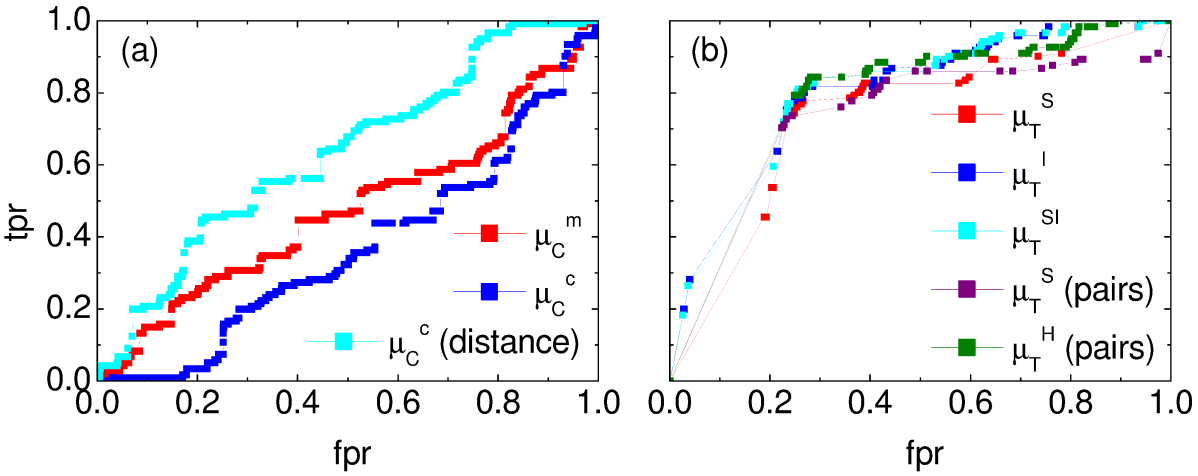

Figure 2.6: (a) ROC curves for the mutual coarse-grained information rate (μ_C^m) and the conditional coarse-grained information rate (interpeted as a similarity μ_C^c (*similarity*) or as a distance measure (μ_C^c (*distance*)), in frames of the identity scoring scheme. (b) Evaluation of the ID scoring scheme using measures based on symbolic dynamics: symbol sequence similarity (μ_T^S), the mutual information of the symbol sequences (μ_T^I) and the mean of these both (μ_T^{SI}), as well as the symbol sequence similarity of pairs of time points (μ_T^S (*pairs*)) and the conditional entropy of the symbols obtained from the pairs of time points (μ_T^H (*pairs*)).

This evaluation leads to the conclusion that only the simple and the residual mutual information can provide a sufficient reconstruction efficiency among the MI measures, if the IDentity scoring scheme is used. However, this is true only if one does not rely on the simple significance test (using reshuffling).

Investigating the performance of the coarse-grained measures on the short gene expression time series, the obtained ROC curves (Fig. 2.6 (a)) are almost the same as expected for a complete random linking in the network. Even though the coarse-grained measures are in principle promising for the inference of interdependency from time series of intermediate length, they are not applicable here. The reason for this is the limited number of time points available (short time series), which makes not only the estimation of the MI, but already the identification of a proper time lag a very challenging task. Interpreting the $CCIR$ as a distance, and not as a similarity measure (as the CMI is assumed to be), leads to an increase of the inferred true positives. However, the predictive power of the measure remains very low.

2.2.3 Model-based measures

The evaluation of the ID scoring scheme using Granger causality as a model-based measure leads to an almost random prediction of links (ROC curves in Fig. 2.5 (b) and (d)). Thus, the Granger causality (GC) measure is not suitable for reconstructing the network, if only very short gene expression time series are available. This is due to the fact that the results of the GC index

depend strongly on the model estimation. Most studies of gene interactions that are based on Granger causality rely on $AR(1)$ processes, which represent only a very vague approximation of the time series. In contrast, to obtain the results shown here, the order of the AR process is determined based on the Akaike information criterion. However, this is insufficient as well, since the AIC usually requires a higher order model (due to the high variability of the data). Hence, the short data are overfitted and the actual interrelations remain covert in most of the cases.

2.2.4 Measures operating on symbolic dynamics

Next, the principle of order patterns is used to derive symbol sequences from the time series [MRTK07]. As already shown in general nonlinear time series analysis, the symbol based measures show in general a good overall performance in reverse engineering.

The ROC curves (Fig. 2.6 (b)) obtained for these measures are rather smooth and flat for fpr's larger than 30%, which means that only a small portion of links is lost when reducing the fpr's down to this value. Consequently, the results are robust to the choice of threshold in that region of the ROC space which is of particular interest when dealing with experimental data.

However, the ROC curves become less smooth for lower values of the fpr's. This implies that fpr's smaller than 20% are barely achievable. The best overall performance has been found here for the combination of symbol sequence similarity and mutual information of the symbol sequences ($SySimMI$), as well as for the mutual information of the symbol sequences ($SyMI$). This outperforms the simple MI of the time series themselves as the length of the series used to estimate the measure is much longer in the case of the symbolic dynamics. Additionally, the conditional entropy of the symbol vectors obtained from pairs of time points shows similar results as the $SySimMI$ and the $SyMI$ in a wide range of the ROC space as well. Furthermore, it shows up that the particular dimension of the order pattern affects the reconstruction only slightly (μ_T^S vs. $\mu_T^S\,(pairs)$).

2.3 Evaluating the reconstruction efficiency

The ROC analysis used above to evaluate the performance of the association measures is a tool for visualizing, organizing, and selecting classifiers based on their performance in terms of a cost/benefit analysis. However, a well-defined rating is not always possible "by eye". Hence, different summary statistics, for example the area under the ROC curve ($AUC(ROC)$) or the $YOUDEN$ index ($YOUDEN = \max(tpr - fpr)$) [FFR05], are common to evaluate and compare the performance of various measures.

Another standard evaluation plot in the field of the ROC analysis is the precision/recall graph ($PvsR$), which is based on the comparison between the true edges and the inferred ones. Thus, it highlights the precision of the reconstruction, and is less affected by the typically large number of false positives in a GRN reconstruction. This renders the area under the precision-recall curve ($AUC(PvsR)$) another plausible summary statistic.

Regarding the $AUC(ROC)$ and the $YOUDEN$ index it becomes apparent (Fig. 2.7 upper and middle panel) that on the short time series, as already suggested by the ROC curves, several

measures perform well in combination with the basic relevance network algorithm: the L^s norm, a few information-theoretic measures (simple and residual mutual information), correlations (simple and conditional Pearson's, as well as Spearman's and Kendall's correlation coefficient) and measures based on symbolic dynamics ($SySim$, $SyMI$ and $SySimMI$). It is obvious that the rank correlations perform better than the Pearson correlation, as well as the mutual information of the symbol sequence works better than the simple MI. This is caused by the fact that symbol and rank based measures are less sensitive to finite size effects and the distribution of data.

While usually the $AUC(ROC)$ and the $YOUDEN$ index show similar trends for the reconstruction efficiency of the various measures, occasionally the $AUC(PvsR)$ reveals additional differences. It shows up that from the previously mentioned well performing measures the L^s norm, Kendall's correlation coefficient and the symbol sequence similarity of pairs are less effective than the other measures. However, the values obtained for the $AUC(PvsR)$ are collectively low here. Hence, the observed trends of the $AUC(PvsR)$ are less meaningful in this case.

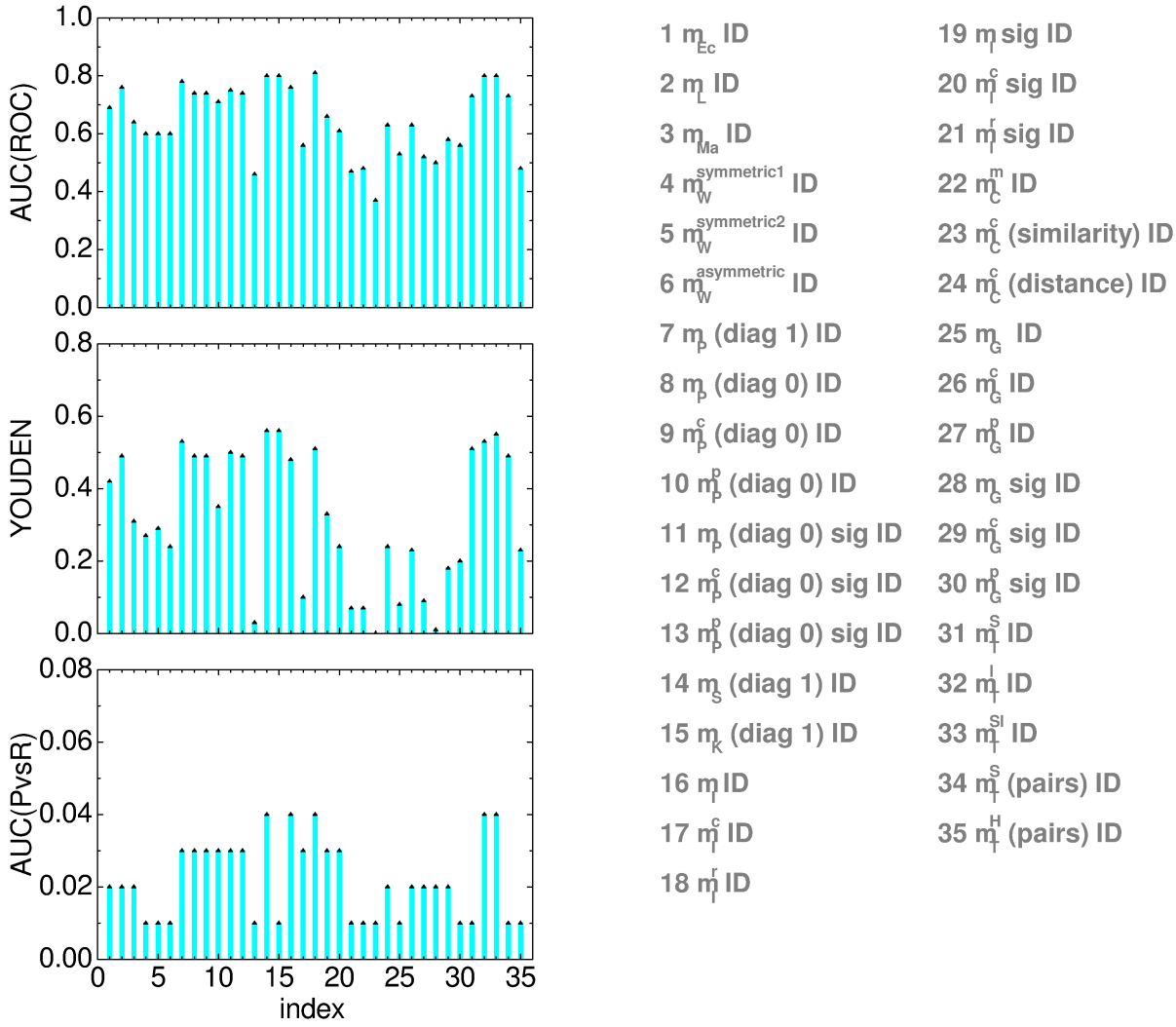

Figure 2.7: Summary statistic of the ROC analysis of various association measures applied for network reconstruction in the framework of the basic relevance network approach. The measures corresponding to the index numbers are given next to the graphs.

3 Evaluating the effect of scoring

Once an association measure has been applied on a given data matrix, there are several possibilities to post process the obtained results by scoring the "weights" of putative interactions. Next, I explain in detail the different scoring schemes (CLR, $ARACNE$, $MRNET$, TS and AWE) which were already mentioned in the previous chapter in Fig. 2.1. In the following, they are applied in Algorithm 1 and there performance on short, synthetic gene expression time series is evaluated. In principle, all of the association measures can be combined with any scoring schemes. The following investigations, however, rely on the most commonly used combinations, and the measures which performed best together with the ID scoring scheme in the previous chapter, respectively.

3.1 Defining scoring schemes

The identity scoring (ID) scheme, employed in Chapter 2, corresponds to the basic relevance network [BK00] approach, where no scoring is involved. However, various other scoring schemes F have been developed, which can be classified into symmetric and asymmetric schemes.

3.1.1 Symmetric scoring

Three strategies, implemented in the R-package "*minet*" (namely CLR, $ARACNE$ and $MRNET$), are applied [Mey09, MLB09]. All of them represent extensions of the basic relevance network approach. They introduce additional scoring rules for the pairwise weighting of the interactions in order to reduce the amount of links that are falsely detected. However, none of these approaches can infer directionality from symmetric association measures.

Context Likelihood of Relatedness (CLR)

An approach often used for the reconstruction of GRN's is the CLR. Weights w_{kl} are assigned to each pair of genes according to the strength of interaction inferred from a particular measure. Then, a score is derived, related to the empirical distribution of the values in W. Thus, the matrix F has the form

$$f_{kl} = \sqrt{\left(\max\left(0, \frac{1}{\sigma_k} - \frac{\bar{w}_k}{w_{kl}\sigma_k}\right)\right)^2 + \left(\max\left(0, \frac{1}{\sigma_l} - \frac{\bar{w}_l}{w_{kl}\sigma_l}\right)\right)^2}, \tag{3.1}$$

where $\bar{w}_k$ ($\bar{w}_l$) and σ_k (σ_l) are the mean and standard deviation of the empirical distribution of w_{km} (w_{lm}), $m = 1, \ldots, q$. The links with $c_{kl} < \tau$ (with $c_{kl} = w_{kl} \cdot f_{kl}$ and τ a predefined

threshold) are removed for the network reconstruction.

The CLR as implemented in R, employes either a **squared correlation matrix** (Pearson's, Spearman's or Kendall's) or the **simple mutual information** to measure the strength of interaction among genes.

Algorithm for the Reconstruction of Accurate Cellular Networks ($ARACNE$)

Furthermore, the Algorithm for the Reconstruction of Accurate Cellular NEtworks, referred to as $ARACNE$, and is included in the current comparison study. The $ARACNE$ is based on the data processing inequality, which states that post-processing cannot increase the amount of information. Hence it follows that:

$$\mu_I(y^{(k)}, y^{(m)}) \leq \min(\mu_I(y^{(k)}, y^{(l)}), \mu_I(y^{(m)}, y^{(l)}), \tag{3.2}$$

when gene k and m are not directly linked, but the coupling is through l, where $y^{(k)}$, $y^{(l)}$ and $y^{(m)}$ are the expression time series of these genes. In this manner, the algorithm discriminates indirect links. First, weights w_{kl} (normalized to the interval $[0, 1]$) are assigned to each pair of nodes. Then the scoring scheme operates as follows: for each triplet of nodes (k, l, m) the edge having the lowest weight (*e.g.*, $w_{k,l}$ in Eq. (3.3)) will be removed, if the difference between the two lowest weights is above a threshold τ_d. In this case, the score f_{kl} is zero and the interaction between k and l is interpreted as indirect. The matrix F obtains the form:

$$f_{kl} = \begin{cases} 0, \text{ if } (w_{kl} \leq w_{lm} \leq w_{mk}) \ \wedge \ (|w_{kl} - w_{lm}| > \tau_d) \\ 1, \text{ otherwise} \end{cases} . \tag{3.3}$$

Moreover, the $ARACNE$ removes all edges satisfying $c_{kl} < \tau$, where τ is a predefined threshold.

To determine the weights the **simple mutual information** or a **squared correlation matrix** is used in the "*minet*"-package. By default, the two thresholds are set to zero.

Maximum Relevance / minimum redundancy NETwork ($MRNET$)

Another example of a advanced relevance network algorithm, is the Maximum Relevance / minimum redundancy NETwork ($MRNET$) [MKLB07]. This scoring scheme performs series of supervised maximum relevance / minimum redundancy ($MRMR$) gene selection procedures, where the expression of each gene k in turn plays the role of the target output $A = y^{(k)}$. Furthermore, the set of the expression data of the input variables of k is $V = y \backslash A$, where y is the set of the expression levels of all genes. Given the set Q of selected variables and pairwise weights $w_{kl|Q}$, the criterion updates Q by choosing the variable

$$y^{(l)}_{MRMR} = arg\ max(s_l),\ y^{(l)} \in V \backslash Q, \tag{3.4}$$

that maximizes the score

$$s_l = u_l - r_l, \tag{3.5}$$

where $r_l = \frac{1}{|Q|} \sum_{m,y^{(m)} \in Q} w_{lm}$ is a redundancy term, and $u_l = w_{lk}$ is a relevance term. Therefore, this scheme assigns higher rank to direct interactions, whereas indirect interactions (redundant information with the direct ones) should receive lower rank. Thus, the matrix F is defined as:

$$f_{kl} = \frac{\max\left[\left(w_{lk} - \frac{1}{|Q|} \sum_{m,y^{(m)} \in Q} w_{lm}\right), \left(w_{kl} - \frac{1}{|Q|} \sum_{m,y^{(m)} \in Q} w_{km}\right)\right]}{w_{kl}}. \tag{3.6}$$

Finally, all edges with a score c_{kl} below a predefined threshold τ are removed.

The implementation of the $MRNET$ in the "*minet*"-package, assigns the weights based on the pairwise **simple mutual information** or a **squared correlation** among the time series of two genes (normalized to the largest value occurring among the pairs).

3.1.2 Asymmetric scoring

Most of the common association measures, such as correlations or the mutual information, are symmetric and cannot distinguish the direction of the interaction. Thus, if they are applied to infer the degree of interaction between pairs of genes, these measures do not allow to draw any conclusion on the drive-response relationships. However, this directionality is very important to correctly infer the regulatory relationship. To extract the probable drive-response relationships from short time-resolved gene expression measurements, two symmetry-breaking scoring schemes for the relevance network approach are proposed here[1].

Time Shift (TS)

In nonlinear time series analysis, the shifting of time series is a common way to infer the directionality of causal relationships. As the driving system has to act first by definition, shifting its time series forward in time (relative to the time series of the response system) should increase the similarity of both time series.

The comparison of the values, which a particular measure obtains for different time delays, suggests the direction of the interaction. Thus, the time shift scoring scheme starts with a cubic spline interpolation for each pair of genes expression time series. Then the series of the second gene is shifted relative to that of the first gene. If x and y are two expression time series stored in the i^{th} and j^{th} row of the data matrix M, and $\tilde{x}$ and $\tilde{y}$ are the related interpolated time series, the shifted time series can be defined as:

$$\tilde{x}_{shift} = \begin{cases} < \tilde{x}_1, ..., \tilde{x}_{N+N_{shift}} >, \text{ if } N_{shift} < 0 \\ < \tilde{x}_{1+N_{shift}}, ..., \tilde{x}_N >, \text{ otherwise} \end{cases} \tag{3.7}$$

[1] Note that the application of symmetry-breaking scoring schemes makes only sense, if a symmetric association measure is used in the relevance network approach.

and

$$\tilde{y}_{shift} = \begin{cases} < \tilde{y}_{1-N_shift}, ..., \tilde{y}_N >, \text{ if } N_{shift} < 0 \\ < \tilde{y}_1, ..., \tilde{y}_{N-N_{shift}} >, \text{ otherwise} \end{cases} \tag{3.8}$$

where N is the length of the interpolated time series and N_{shift} is the assumed shift of $\tilde{y}$ versus $\tilde{x}$, with $N_{shift} \in \mathbf{Z}$ and $N_{shift} \in [-0.1 \cdot N, 0.1 \cdot N]$. Next, $\mu(\tilde{x}_{shift}, \tilde{y}_{shift})$ is evaluated for all possible values of N_{shift}, resulting in a vector $< \mu_{-N_{shift}}, ..., \mu_{N_{shift}} >$ (for not significant values of μ the corresponding entry will be set equal 0). The scoring is given by

$$f_{kl} = \begin{cases} 1, \text{ if } \max\left[< \mu_{-N_{shift}}, ..., \mu_0 >\right] \geq \max\left[< \mu_1, ..., \mu_{N_{shift}} >\right] \\ 0, \text{ otherwise} \end{cases} . \tag{3.9}$$

If the largest significant value of the measure is obtained for a negative shift, the regulatory direction from the first to the second gene is kept, while the opposite direction is preserved if the largest significant value is obtained for a positive shift. Furthermore, both regulatory directions are kept, if the maximum arises for a shift of zero or multiple opposed shift values or in the case when no significant value exists. The aim of the scoring scheme is to suggest a direction. However, the scheme does not rely on the absolute values of the correlations which are calculated from the delayed time series, because these values are rather biased as the data sets are quite short.

In the next step of Algorithm 1, the information regarding the directionality are combined with the weight of interaction inferred from a particular association measure ($c_{kl} = w_{kl} \cdot f_{kl}$). The weights for the unlikely direction are set to zero in order to break symmetries, and thus reduce the number of false positive links.

Finally, all edges with $c_{kl} < \tau$ are removed, where τ is a particular threshold.

The TS scoring scheme is tested using the **absolute value of the correlation coefficients** μ_P **(Pearson)** and μ_S **(Spearman)** for pairs of the shifted expression series, where the significance level was set to $\alpha = 0.01$ and only absolute values of correlation larger 0.9 have been taken into account. The measure to infer the weights in the first step of Algorithm 1 is either the **mean of sequence similarity and mutual information of symbols** or **Spearman's** and **Pearson's correlation**.

Furthermore, this scoring scheme is applied in addition to (or after) another scoring scheme (*e.g.*, ID, CLR or AWE). It is important to note that in contrast to the previously described modifications of the algorithm, the proposed scoring scheme allows to investigate the directionality, when symmetric association measures are considered.

A novel scoring scheme – Asymmetric WEighting (AWE)

In the following, I introduce a **novel** asymmetric weighting based on topological aspects. This weighting approach is applied to the complete set of pairwise weights obtained from a particular association measure, and it is implemented according to Algorithm 1. In particular, a matrix of weights is computed, where the columns represent the genes which are regulated, and the rows the ones which regulate other genes. The scoring value is then calculated by dividing

each row entry by the sum of the corresponding column values. The scoring scheme (and the corresponding matrix F) is defined as:

$$f_{kl} = \left(\sum_{m=1}^{q} w_{ml} \right)^{-1}. \tag{3.10}$$

Hence, the probability that the l^{th} gene is regulated sums up to one:

$$\sum_{m=1}^{q} c_{ml} = \sum_{m=1}^{q} (w_{ml} \cdot f_{ml}) = 1. \tag{3.11}$$

The score indicates how likely a gene is regulating another one. Here, this value depends not only on the strength of interactions, it also depends on the amount of inputs.

Eventually, if $c_{kl} \geq \tau$ the edge is introduced, otherwise it is omitted.

The asymmetric weighting is tested on the matrix W inferred from the **symbolic dynamics measures**.

3.2 Performance of various scoring schemes in terms of receiver operating characteristics curves

In order to further investigate the reconstruction efficiency of the generalized relevance network approach, I compare the performance of the different scoring schemes. I use ROC curves to evaluate the performance, as it was done in the previous chapter for the association measures. The evaluation is again based on short, synthetic gene expression time series (10 time points, no noise) of the network of 100 genes of *E. coli*.

3.2.1 Symmetric scoring schemes

First, I discuss the reconstruction efficiencies of the three symmetric modifications of the relevance network algorithm which are defined above. For this purpose the CLR, the $ARCANE$ and the $MRNET$ as implemented in the "*minet*"-package are applied, where the default weight of the pairwise association measure, namely the squared Spearman's correlation for every set of pairs, is used. The corresponding results are presented in Fig. 3.1 (a), in terms of the ROC curves.

Since the extended relevance network algorithms implemented in the "*minet*" are designed to reduce the number of false positives, high fpr's (of more than about 50%) do not occur here, unless all interactions are set as links.

Moreover, the $MRNET$ and the CLR result in ROC curves which are not smooth, meaning that their capability to reconstruct particular links is limited and strongly dependent on a proper choice of the threshold τ. The $ARACNE$, on the other hand, is restricted to an almost fixed fpr-tpr value.

However, none of these scoring scheme is able to indicate directionality from symmetric measures.

3.2.2 Asymmetric scoring schemes

In the following paragraph, I investigate the performance of the Time Shift (TS) as a symmetry-breaking scoring scheme (apparently for the first time in GRN reconstruction). The results of this modification of the relevance network algorithm are presented and evaluated in the cases of

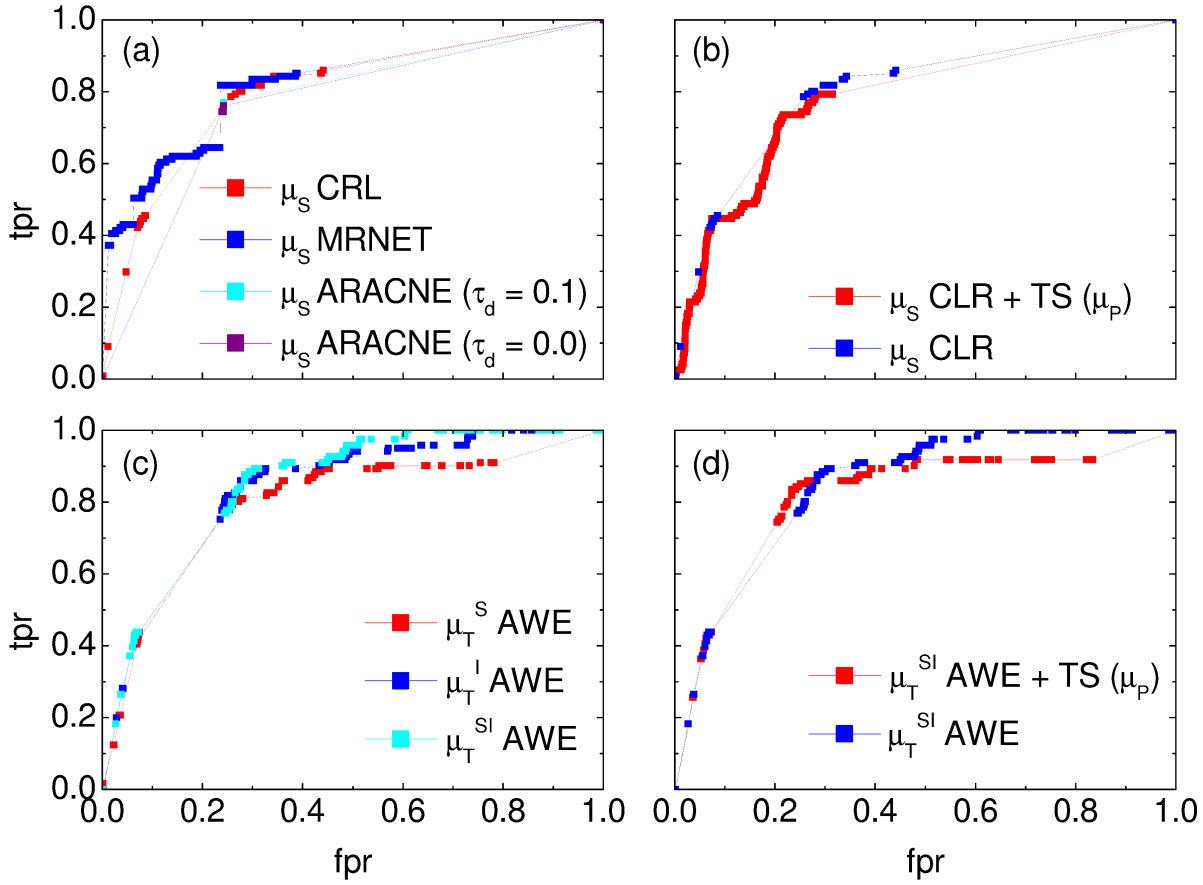

Figure 3.1: (a) ROC curves obtained for the Spearman correlation coefficient μ_S using the CLR, $MRNET$ and the $ARACNE$ scoring scheme. (b) The corresponding ROC curves illustrating the performance of the TS scoring scheme using the Pearson correlation μ_P, applied in addition to the CLR (measure: μ_S). (c) Performance of the AWE algorithm using the selected symbol based measures included in the this study, for example ROC curves for the symbol sequence similarity (μ_T^S), the mutual information of the symbol sequences (μ_T^I), and the mean of these both (μ_T^{SI}). (d) The ROC curves when TS is applied in addition to the AWE (measure: μ_T^{SI}) scoring scheme.

removing the links which are falsely detected by the CLR (measure: μ_ρ) or the AWE (measure: μ_T^{SI}). However, unraveling the directionality of interaction between pairs of genes using the correlation of the delayed time series decreases the maximal achievable tpr's here.

The slope of the ROC curves (shown in Fig. 3.1 (b)) indeed does not differ much from the results of the CLR and AWE scoring schemes. Moreover, if the TS scoring scheme is based on Pearson's correlation, the ROC curve obtained from the $CLR + TS$ is considerably smoother than the curve where TS is not applied. Hence, the prediction is less sensitive to the choice of a threshold. The same is not true for the ROC curve obtained when applying the TS scoring scheme in addition to the AWE (Fig. 3.1 (d)).

Instead, this curve becomes more flat and is slightly shifted towards lower fpr's by contrast to the corresponding curve in Fig. 3.1 (d) where TS is not applied. That means, while for low fpr the curve looks basicly the same, in the intermediate range of the ROC space (fpr of about 0.15 to 0.45) similar tpr values can be obtained for lower fpr. However, in the range of high fpr the maximal achievable tpr value is lower. Hence, tpr's of approximately 80% can be achieved with lower costs, as the according number of false positives is in general smaller, if the TS scoring scheme is used. On the other hand, as already mentioned, the quality of the link detection becomes worse for higher fpr's (more than about 40%) compared to the corresponding results of the AWE itself. The true positives in the ROC curve in Fig. 3.1 (d) are almost constant in that region of the ROC space.

As well as the TS, the AWE scoring scheme aims at breaking symmetries and thus, allowing extraction of information about the directionality of interaction from symmetric association measures. However, a detailed comparison of the reconstruction efficiency of the AWE using different symbolic dynamics measures shows that in contrast to the TS scoring scheme, AWE does not decrease the maximal achievable tpr's.

The ROC curves shown in Fig. 3.1 (c) are flatter for high fpr compared to the curves obtained for the basic algorithm with the ID scoring scheme using the same symbolic dynamics measure (Fig. 2.6 (b)). Hence, tpr's of more than 80% are achievable by the AWE algorithm with much lower costs than with the ID scoring scheme. On the other hand, the ROC curves obtained from AWE are more steep for low fpr's. This implies that tpr's up to approximately 45% can be achieved with fpr's of less than 10% here. Furthermore, the curves shown in Fig. 3.1 (c) are much smoother in comparison to those in Fig. 2.6 (b), indicating that the reconstruction is less sensitive to the choice of a particular threshold.

3.3 Ranking of association measures and scoring schemes

To evaluate and rank the overall performance of all approaches under study the three common summary statistics of ROC analysis that have been already applied in Section 2.3, namely the area under the ROC curve ($AUC(ROC)$), the $Youden$ index and the area under the Precision/Recall curve ($AUC(PvsR)$), are employed. Furthermore, as the modifications of the algorithm implemented in the "minet" package (CLR, $MRNET$ and $ARACNE$) are commonly and widely used as approaches for GRN reconstruction they serve as a benchmark for the comparison of the different measures and scoring schemes. In order to determine proper benchmarks the

summary statistics are evaluated for the three algorithms together with the different measures, and estimators thereof respectively, which are implemented in the "minet" package.

In Tab. 3.1 an overview of these results is given, leading to the following benchmarks: a measure combined with a particular scoring scheme is called

- well performing for short expression data sets (evaluated on the synthetic data in this case) if
  $AUC(ROC) > 0.8$,
  $YOUDEN > 0.5$ and
  $AUC(PvsR) > 0.05$,
- sufficiently performing if
  $0.8 > AUC(ROC) > 0.7$,
  $0.5 > YOUDEN > 0.4$ and
  $0.05 > AUC(PvsR) > 0.03$,
- and deficient otherwise.

The ROC summary statistics for 50 combinations of association measures and scoring schemes are shown in Fig. 3.2.

The modifications of the relevance network algorithm in the "*minet*" package having the best performance in the reconstruction of GRN from short data sets, are the CLR and the $MRNET$ ("*minet*" is based on Spearman's correlation in this case). Here the $AUC(ROC)$ indicates almost no change compared to the basic algorithm with identity scoring (measure: Spearman's correlation), while the $YOUDEN$ index decreases for the CLR and increases for the $MRNET$. However, the opposite is true for the $AUC(PvsR)$. The overall performance of the CLR (in terms of the considered summary statistics) is slightly better than those of the $MRNET$ (CLR scoring scheme was used to set the benchmarks).

In combination with ID scoring scheme several measures based on random variable (the simple and residual MI, simple and conditional Pearson's as well as Spearman's correlation coefficient) and measures based on symbols ($SyMI$ and $SySimMI$) perform sufficiently well. Additionally, the L^s norm, Kendall's rank correlation and the symbol sequence similarity perform also well with respect to $AUC(ROC)$ and $YOUDEN$.

Several of these well performing measures are further applied together with the asymmetric scoring scheme. In such cases, the measures combined with the Time Shift scoring scheme perform sufficiently well, however, the summary statistics do not change much compared to the results obtained for the same measures using the ID. In contrast to this, the asymmetric weighting yields a significant increase among all the summary statistics compared to the performance of the same measures using only the ID scoring scheme.

Finally, the approaches with the highest capability to detect true, and eliminate false positive links at the same time are ranked as follows:

- μ_T^{SI} $AWE + TS$ (scoring by μ_S),
- μ_T^{SI} $AWE + TS$ (scoring by μ_P),

parameter (minet)	noiselevel 0.0		
	AUC(ROC)	YOUDEN	AUC(PvsR)
clr, mi.empirical, equalfreq	0.80	0.54	0.05
clr, mi.empirical, equalwidth	0.76	0.45	0.04
clr, mi.mm, equalfreq	0.80	0.54	0.05
clr, mi.mm, equalwidth	0.76	0.48	0.04
clr, mi.shrink, equalfreq	0.80	0.53	0.05
clr, mi.shrink, equalwidth	0.74	0.41	0.04
clr, mi.sg, equalfreq	0.80	0.54	0.05
clr, mi.sg, equalwidth	0.74	0.42	0.04
clr, pearson, none	0.78	0.49	0.05
clr, spearman, none	0.80	0.53	0.05
clr, kendall, none	0.80	0.53	0.05
mrnet, mi.empirical, equalfreq	0.82	0.59	0.04
mrnet, mi.empirical, equalwidth	0.76	0.47	0.05
mrnet, mi.mm, equalfreq	0.81	0.57	0.04
mrnet, mi.mm, equalwidth	0.77	0.46	0.05
mrnet, mi.shrink, equalfreq	0.81	0.57	0.04
mrnet, mi.shrink, equalwidth	0.73	0.39	0.04
mrnet, mi.sg, equalfreq	0.81	0.57	0.04
mrnet, mi.sg, equalwidth	0.77	0.47	0.06
mrnet, pearson, none	0.78	0.49	0.04
mrnet, spearman, none	0.82	0.58	0.03
mrnet, kendall, none	0.81	0.56	0.03
aracne, mi.empirical, equalfreq	0.76	0.52	0.01
aracne, mi.empirical, equalwidth	0.54	0.12	0.02
aracne, mi.mm, equalfreq	0.76	0.52	0.01
aracne, mi.mm, equalwidth	0.54	0.12	0.02
aracne, mi.shrink, equalfreq	0.76	0.52	0.01
aracne, mi.shrink, equalwidth	0.55	0.14	0.02
aracne, mi.sg, equalfreq	0.76	0.52	0.01
aracne, mi.sg, equalwidth	0.54	0.12	0.02
aracne, pearson, none	0.54	0.07	0.03
aracne, spearman, none	0.76	0.52	0.01
aracne, kendall, none	0.76	0.52	0.01

Table 3.1: Overview of the summary statistics from the ROC analysis for different algorithms, association measures and estimator implemented in the minet package.

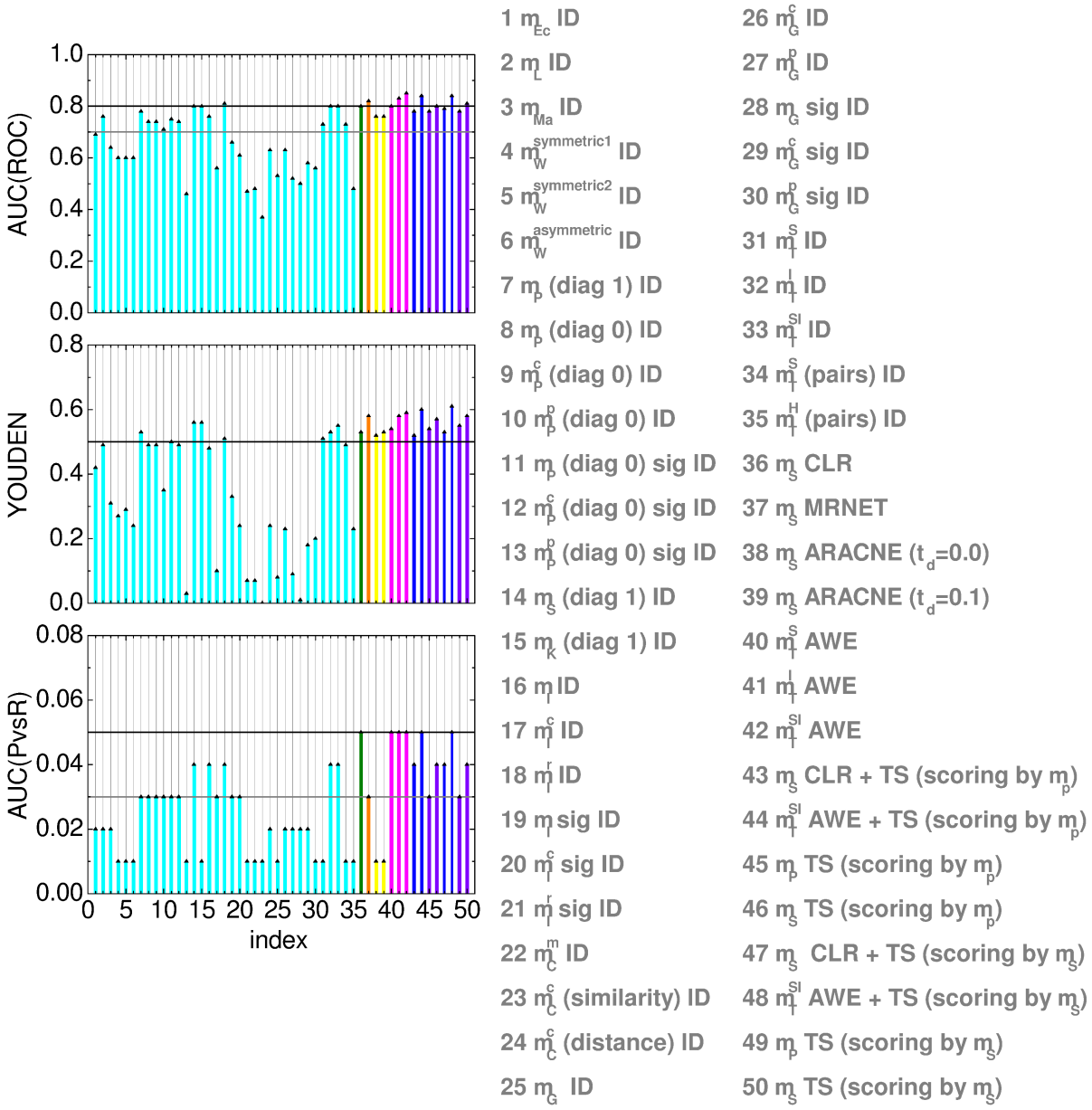

Figure 3.2: Evaluation of the investigated scoring schemes and measures using three summary statistics from ROC analysis. Similar approaches are grouped together. The first group in cyan refers to the different association measures applied together with the ID scoring scheme. The green stands for the CLR, orange for the $MRNET$ scoring scheme. Yellow refers to the $ARACNE$, magenta to the AWE and violet to the TS scoring scheme. These colors are related to those in Fig. 2.1. Furthermore, blue groups together all measures applied with a combination of scoring schemes. The indices refer to the measures and scoring schemes listed next to the graphic.

- μ_T^{SI} AWE,
- μ_T^{I} AWE,
- μ_T^{S} AWE and
- μ_S CLR.

The asymmetric weighting (AWE) significantly improves the prediction at this point, since it breaks the symmetry of a particular measure based on topological considerations, and therefore reduces the number of false positive links. Hence the AWE (measure: μ_T^{SI}) clearly shows the best performance when short time series are considered (the results become slightly better if Time Shift is applied in addition).

4 Influences on the reconstruction efficiency

Reverse engineering of GRN's represents one of the crucial topics in contemporary systems biology and bioinformatics research. However, the large number of genes interacting in a complex manner versus the short, coarsely and sometimes irregularly sampled, and noisy expression time series which are usually available, renders the reconstruction a challenging task. Hence, I investigate separately the role that noise, sampling and interpolation, as well as the size and topology of the network play for the ability to correctly infer the links from time-resolved data.

4.1 The role of noise

In general, noise-free expression measurements cannot be achieved in real experiments. In fact, intermediate and high noise levels are not rare. In order to account for stochasticity in the time series, and to establish the robustness of the ranking of the investigated association measures and scoring schemes, next, I evaluate their performance for noisy synthetic time series obtained for the same network of 100 genes of *E. coli* as in the previous chapters.

First, synthetic time-resolved gene expression data at noise level 0.3 are considered. As expected, the measures which failed in the noise-free case (*e.g.*, DTW, CMI, the coarse-grained information rate and the Granger causality measures) did not improve their performance. Fig. 4.1 shows the corresponding ROC curves. On the other hand, it also shows up that the measures based on vectors, produce very robust results with respect to noise. However, since the performance of these measures was already insufficient in the noise-free case, its general overall ranking does not improve significantly. It must be further noted that the measures which performed best in the case of noise level 0.0, such as MI, RMI, correlations and symbol based measures differ in their robustness with respect to noise, as illustrated in Fig. 4.2.

For example, the reconstruction efficiency of the simple and the conditional Pearson's correlation slightly decreases, while that of partial Pearson's correlation slightly increases. Hence, all three measures result basically in the same ROC curves, *i.e.*, the computationally intensive calculation of partial and conditional Pearson's correlation can be avoided under these circumstances. Furthermore, MI and RMI both lose their accuracy as noise increases, and the corresponding ROC curves resemble those of the Pearson's correlation. However, the relation between both measures stays the same: RMI performs only slightly better than MI which very likely results from the poor reliability of the estimation of the propability distribution from the limited data.

Moreover, the reconstruction efficiency for the symbol based measures decreases significantlyl, which is true in particular for the mutual information of symbol sequences. Apart from that, the ROC curves obtained for the symbol based measures are more continuous for noisy data than those in the noise-free case. This implies that the reconstruction process in that case is less governed by the choice of a particular threshold.

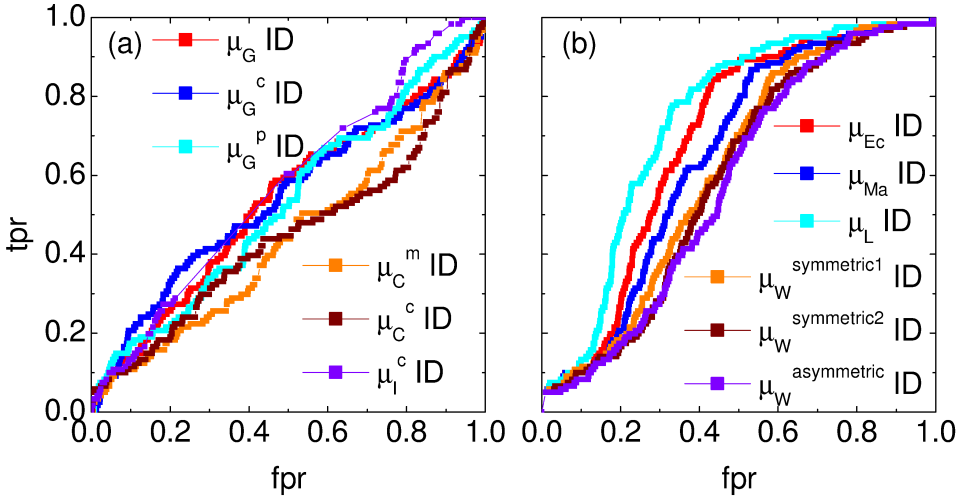

Figure 4.1: Reconstruction from noisy data (noise level 0.3). ROC curves of (a) the Granger, partial Granger and conditional Granger causality (μ_G, μ_G^p, μ_G^c), the mutual and conditional coarse-grained information rates (μ_C^m, μ_C^c), and the conditional mutual information (μ_I^c), as well as (b) the distance measures: L^s norm (μ_L), Euclidean distance (μ_{Ec}), Manhattan distance (μ_{Ma}) and dynamic time warping (μ_W) with the step pattern symmetric1, symmetric2 and asymmetric.

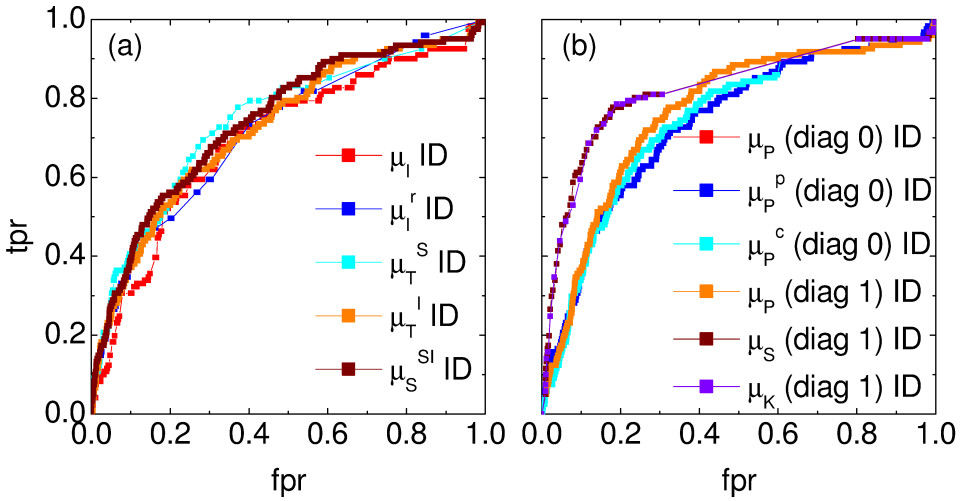

Figure 4.2: Performance of various similarity measures for noisy data (noise level 0.3). The plot shows ROC curves of (a) mutual information (μ_I), residual mutual information (μ_I^r), symbol sequence similarity (μ_T^S), mutual information of the symbol sequences (μ_T^I) and the mean of these two (μ_T^{SI}), and (b) Pearson correlation (μ_P), partial Pearson correlation (μ_P^p), conditional Pearson correlation (μ_P^c), Spearman correlation (μ_S) and Kendall correlation (μ_K).

A similar behavior is observed for the rank correlation coefficients. However, the shape of the curves appears more robust under the influence of noise than the ones for the symbol based measures. Thus, the rank based measures represent the most suitable association measures to study the interrelation among short time series for increasing noise levels.

Finally, it has to be mentioned that the CLR and the AWE are the most robust scoring schemes with respect to noise, whereas $ARACNE$ fails for short and noisy time series. $MRNET$ and TS show intermediate dependence on the noise intensity.

Next, to establish the robustness of the investigated top-ranked association measure against noise, their performance in terms of ROC statistics is evaluated for two different noise intensities, namely 0.3 (Fig. 4.3) and 0.5 (Fig. 4.4). Only measures which perform sufficiently well in the noise-free case (measures operating on random variables and symbolic dynamics) are tested. In particular, Pearson's (μ_P), Spearman's (μ_S) and Kendall's (μ_K) correlation coefficients as well as the symbol based measures (μ_T^S, μ_T^I, μ_T^{SI}, and μ_T^H) are examined using the ID scoring scheme. In addition, the performance of CLR, $MRNET$, $ARACNE$, AWE and TS scoring schemes is investigated based on the same measures as in the noise-free case.

As already suggested by the ROC curves under the influence of noise the quality of the results of the symbol based measures (in particular μ_T^I) decreases. Noise strongly influences the process of symbol assigning (as well as the binning process for MI calculation), and thus, it can principally enhance or distort the information content. The direction of the influence is not predictable *a priori*, but in the presence of strong noise, symbols are no longer reliable (if no additional information on the influence of the noise is provided). On the other hand, measures operating on random variables are rather robust against noise. The best results in these cases have been achieved using rank correlations.

The $ARACNE$ has proven once more to be very sensitive with respect to noise. In contrast to this, the AWE (compared to the results of the ID scoring using the same measure) still performs well within the given limits, as it is only based on topological considerations, and it is not influenced by the presence of noise.

4.1.1 Influence of the length of the time series

To further substantiate the obtained results on the robustness of the network reconstruction under noisy conditions, the area under the ROC curve and the $YOUDEN$ index are calculated and depicted in Fig. 4.5 as a function of the noise intensity. The 5 combinations of association measures and scoring schemes which performed best in the noise-free case, namely the symbolic measures and the asymmetric scoring schemes, mentioned previously, are considered. Additionally, I compare these results to those obtained for time series of different lengths (*i.e.*, 8 and 20 time points).

The evaluation of the area under ROC curve and the $YOUDEN$ index prove that for short time series the capability of the measures and scoring schemes to detect true and at the same time eliminate false positive links depends both on the number of time points and the noise intensity. However, this dependence is small compared to the differences which were observed in Section 3.3 in the reconstruction efficiency between the different association measures and scoring schemes.

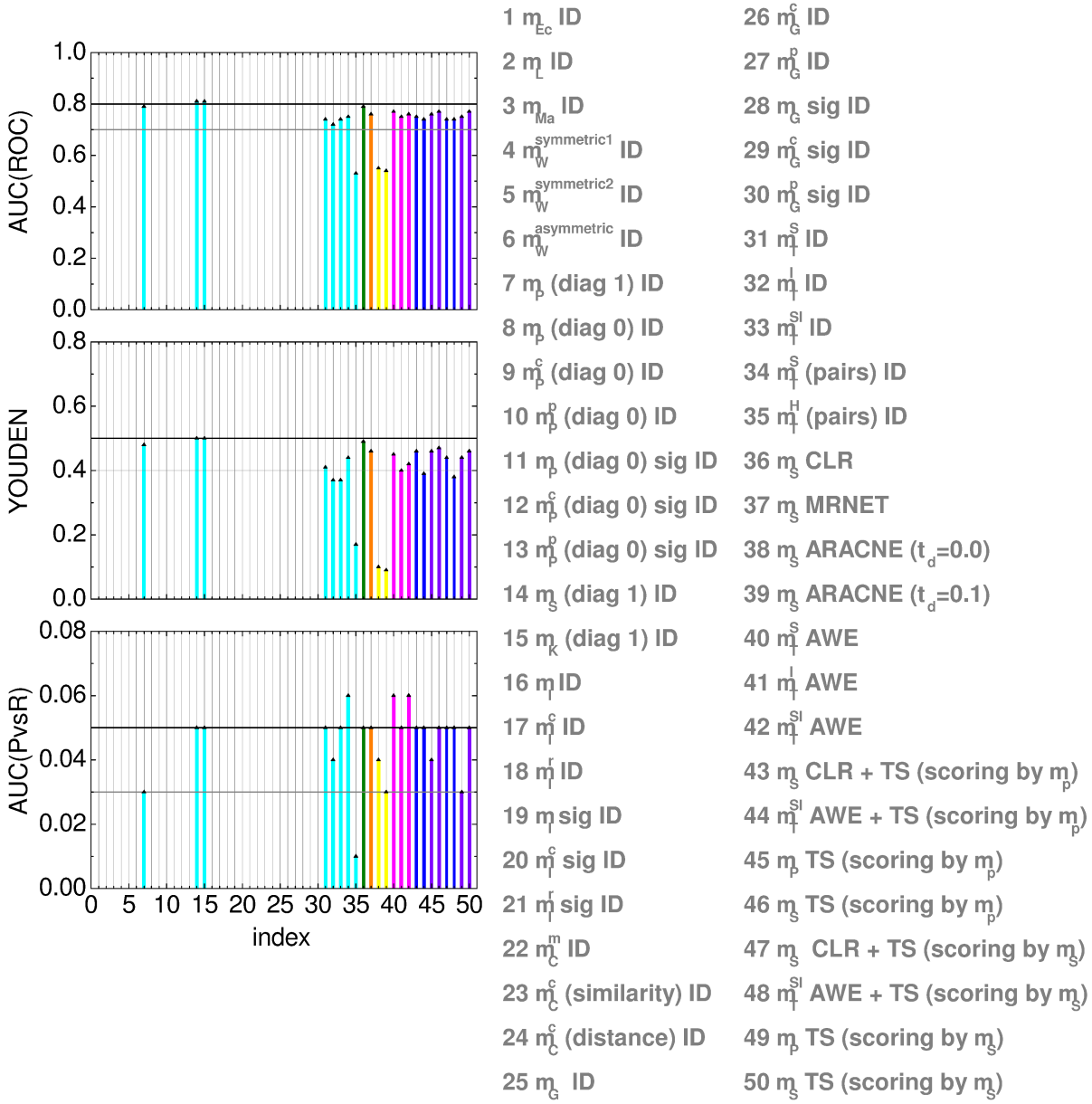

Figure 4.3: Summary statistics for the top-ranked measures / scoring schemes for increased noise intensity (noise level 0.3). Similar approaches are grouped together. The first group in cyan refers to the different measures applied together with the ID scoring scheme. The green stands for the CLR scoring scheme, the orange for the $MRNET$, yellow refers to the $ARACNE$, magenta to the AWE and violet stands for the TS. Furthermore, blue groups together all measures applied with a combination of scoring schemes. The indices refer to the measures and scoring schemes listed next to the graphic.

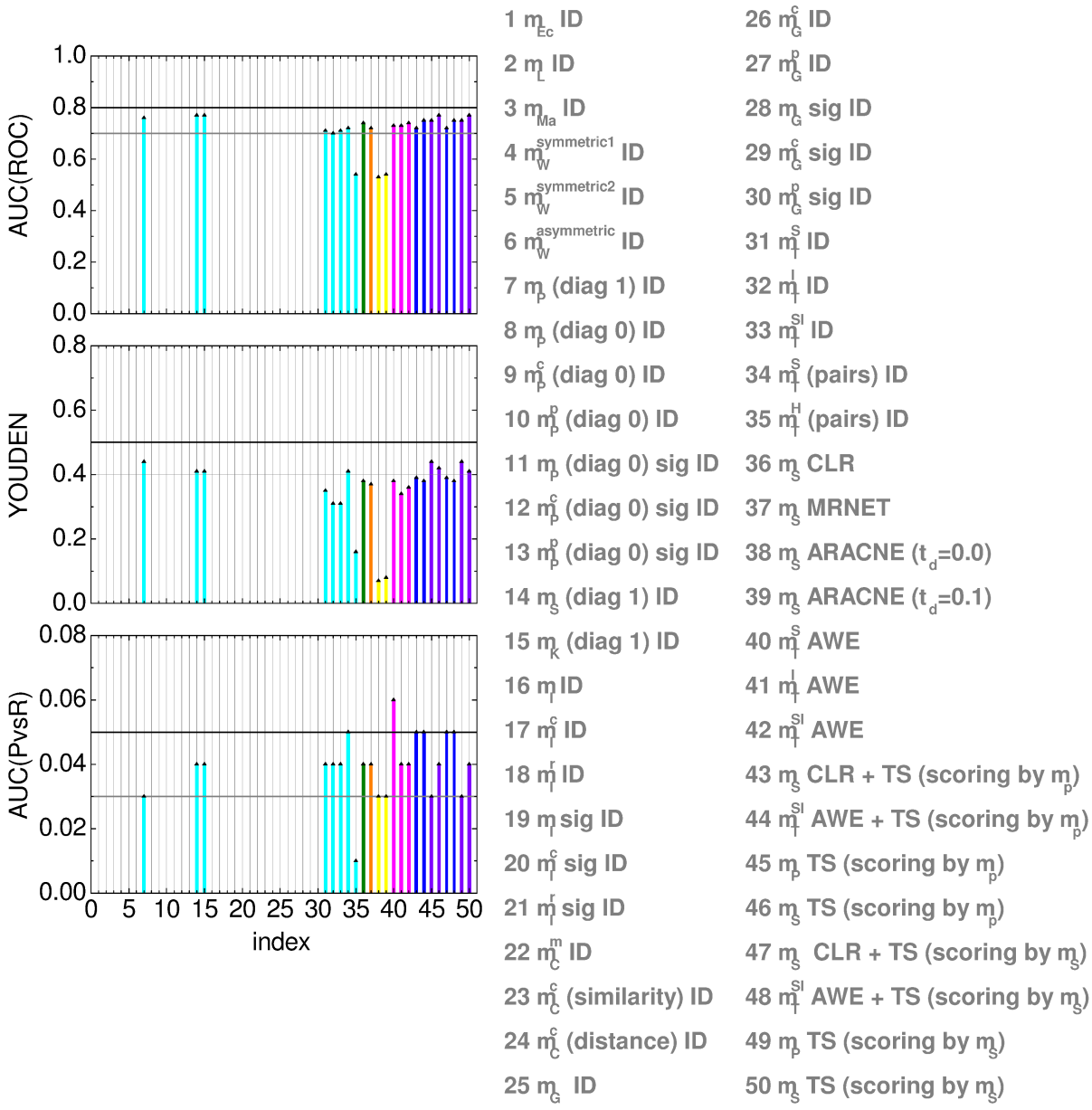

Figure 4.4: Summary statistics for the top-ranked measures / scoring schemes for noise level 0.5. Color refer to the same groups as in Fig. 4.3

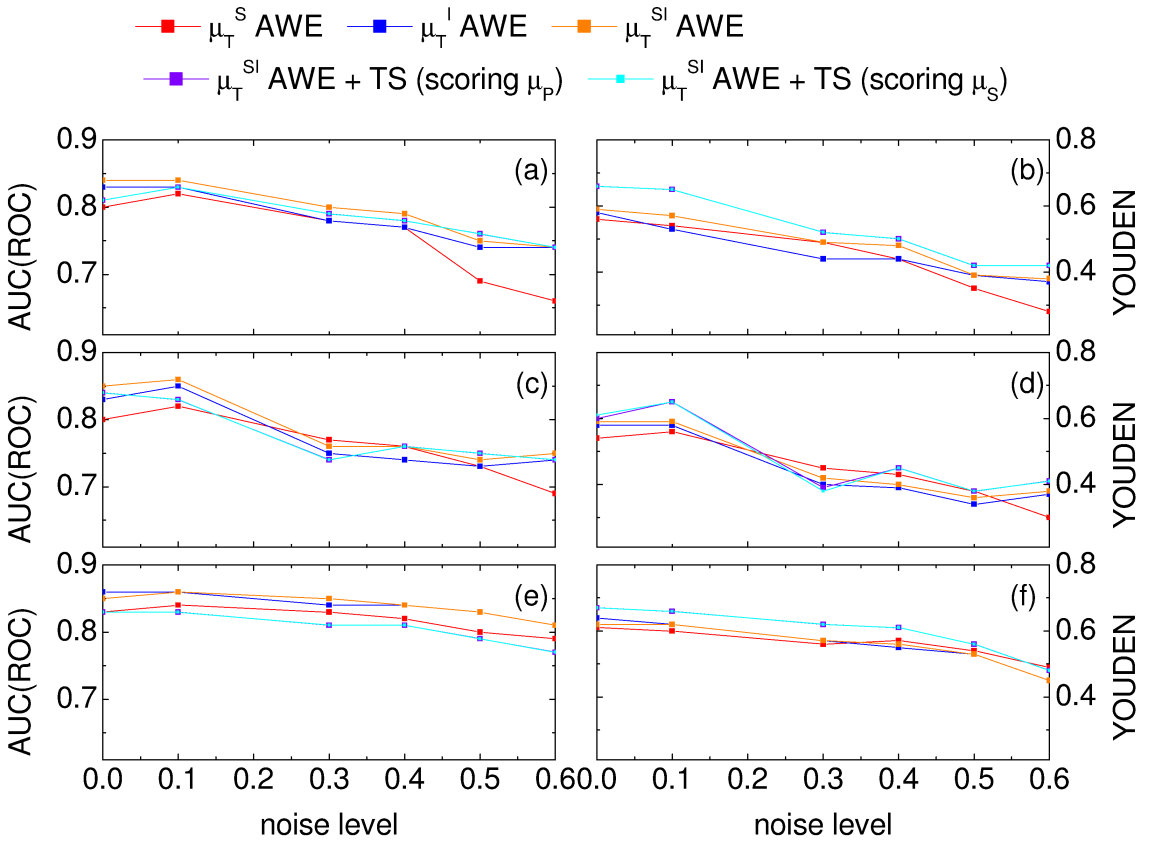

Figure 4.5: Summary statistics ((a), (c) and (e) area under the ROC curve, as well as (b), (d) and (f) *YOUDEN* index) for the top-ranked measures / scoring schemes as a function of the noise intensity for varying lengths of the time series. The results in (a) and (b) are obtained for 8 time points, those in (c) and (d) for 10 time points, and those in (e) and (f) for 20 time points.

The sensitivity with respect to noise is reduced if the length of the time series is increased (which corresponds to the usage of order pattern of higher dimension). Moreover, in general, the reconstruction efficiency decreases if the noise level increases or the length of the time series decreases. For the short time series I used in this study, however, these dependencies are not monotone.

4.2 The role of interpolation and sampling

Due to the fact that time-resolved gene expression data are usually quite coarsely sampled, assured assumptions upon what happens between two time points cannot be achieved in general. This problem becomes obvious when unequally sampled data are used (shown in Fig. 4.6 (a) for synthetic gene expression time series of length 10). Although the interpolation at the beginning of the time series (where the time points are rather close) seems to be sufficient, it does not capture the dynamics of the simulated expression time series any longer when the distance between the time points becomes larger. Hence, by interpolating the gene expression data sets, artifacts are introduced, which will be further reflected in the results of the particular association measures. In order to avoid these artifacts interpolation is usually not applied in this study, even though this leads to less significant results for almost all measures, as they operate far below the limit of their theoretically defined preconditions. Nevertheless, the overall results

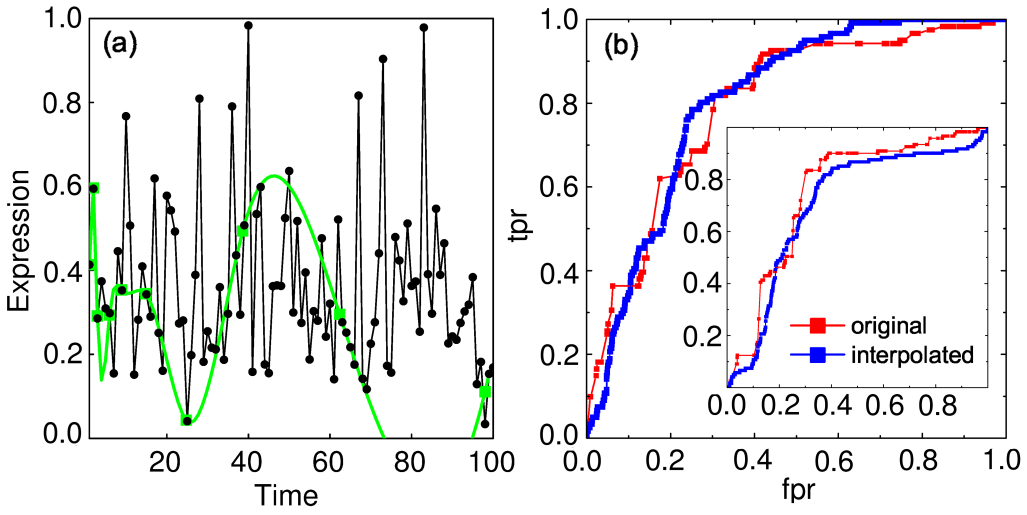

Figure 4.6: (a) Simulated expression time series of 100 equally sampled data points (black line), and the effect of (spline) interpolation using the following data points of the original series: 1|2|3|6|9|15|25|39|63|99 (green line). (b) Artefacts introduced in the reconstruction procedure (measure: μ_I, scoring scheme: ID) by interpolation of short, coarsely sampled time series. ROC curves are shown for the noise-free case and 10 time points equally sampled in time, and unequally sampled (inset plot), respectively.

(ROC analysis) have been observed to be very similar or even better when interpolation is not included, especially when non-uniformly sampled time series are considered. Fig. 4.6 (b) illustrates this effect exemplary for the simple mutual information.

However, some measures, such as the Granger causality, as well as several scoring schemes (*e.g*, the TS), are explicitly time dependent. Hence, they require uniformly sampled data, *i.e.*, an interpolation is needed if only non-uniformly sampled data is available which is frequently the case when reconstructing GRN's.

On the other hand, most of the well performing reconstruction methods in this study are not explicitly time-dependent and do not require a specific time sampling. This implicates that they are not very sensitive concerning the spacing on the time axis. Additionally, Fig. 4.6 (b) illustrates that a non-uniform sampling for these methods can even improve the quality of the reconstruction, as a larger period of the dynamics is captured. This is further confirmed by comparing the results in Fig. 4.7 to those shown in the previous sections.

4.3 The role of the network topology

In general, the underlying network and its properties are not known prior to the reconstruction process. This renders a case by case study of topological effects on the reconstruction efficiency impossible due to the immense amount of imaginable topologies. However, the available experimental and theoretical research has suggested that GRN's most likely are characterized with scale-free properties [Alb05], as discussed in Section 1.2.

Therefore, I investigate the reconstruction efficiency of the relevance network approach for various subnetworks of *E. coli* and *S. cerevisiae*. The presented results are obtained for subnetworks of distinct sizes which differ in degree and clustering coefficient (Fig. 4.8). In particular, the *E. coli* subnetwork of 100 genes, 10 of which code for transcription factors, as used in the previous chapters, is revisited. This network includes 121 links and is characterized by an average degree of 2.42 and a clustering coefficient of 0.016. The obtained results are compared to those for the following two networks: (i) An *E. coli* subnetwork of 200 genes (34 coding for transcription factors) that includes 303 links and is characterized by an average degree of 3.03 and a clustering coefficient of 0.019, and (ii) a *S. cerevisiae* subnetwork of 100 genes (14 coding for transcription factors) that includes 123 links and is characterized by an average degree of 2.46 and a clustering coefficient of 0.026.

Although the three networks differ in several of their properties, the performance of the top-ranking measures, such as the symbol based measures, rank correlations, MI and RMI is not much affected: very similar ROC curves are obtained for all of the network types analyzed, as shown in Fig. 4.9 (this also pertains for several other measures, as shown in Fig. 4.10).

The performance in the range of low fpr is improved for most of the measures for increased average degree of nodes. However, at the same time the performance in the range of high fpr is usually decreased.

In general, the largest differences in the reconstruction efficiency occur for the conditional Granger causality and partial Pearson correlation (Fig. 4.9 (a) and (b)), where the quality of the reconstruction decreases significantly for an increased number of nodes (*e.g.*, *E. coli* network of 200 nodes) and an increased clustering coefficient, as in the *S. cerevisae* network.

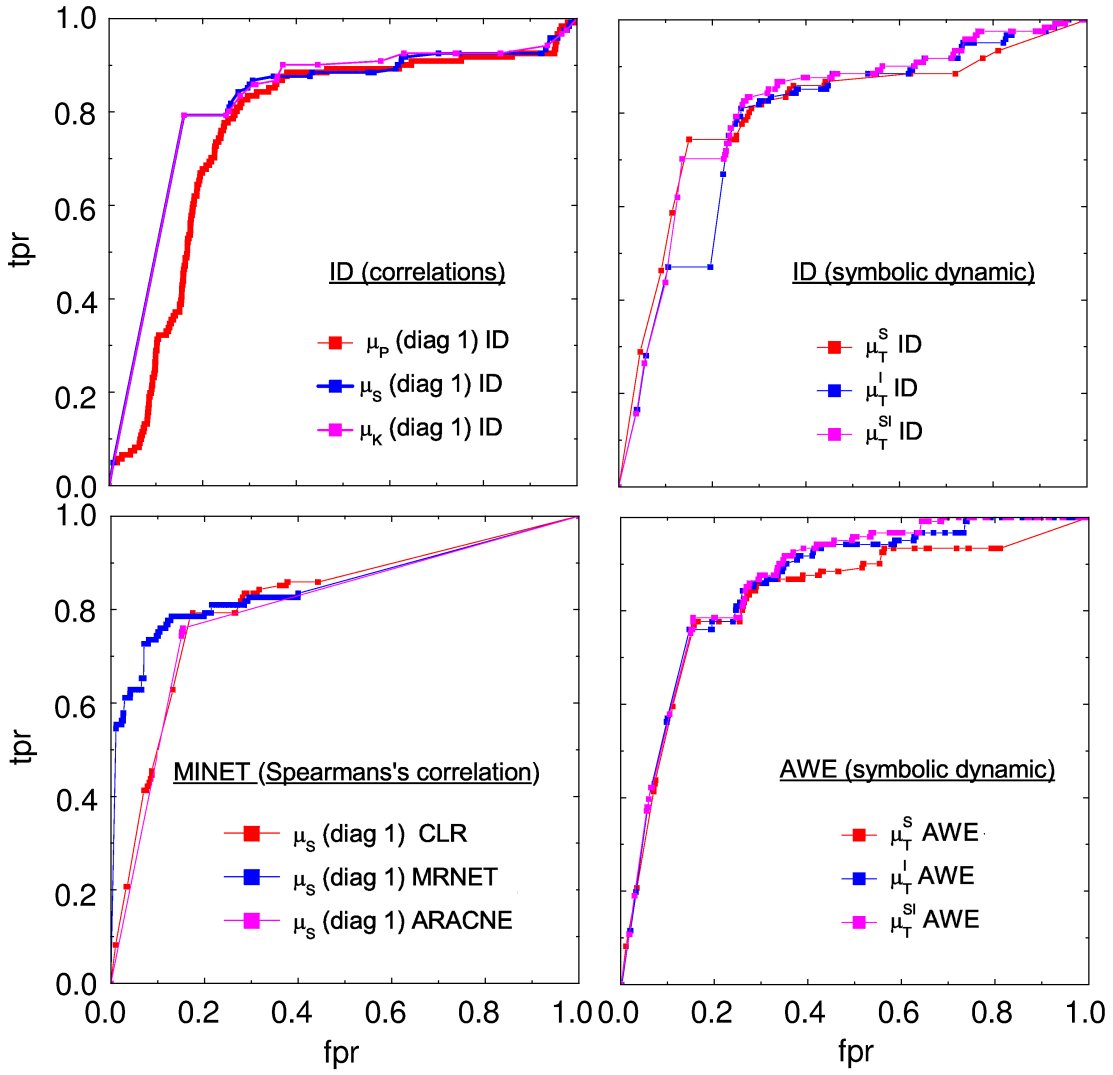

Figure 4.7: ROC curves for selected measures and algorithms obtained in the noise-free case, using unequally sampled data without interpolation. The sampling is the same as in the previous figure, including the following data points of a simulated series of 100 points: 1|2|3|6|9|15|25|39|63|99.

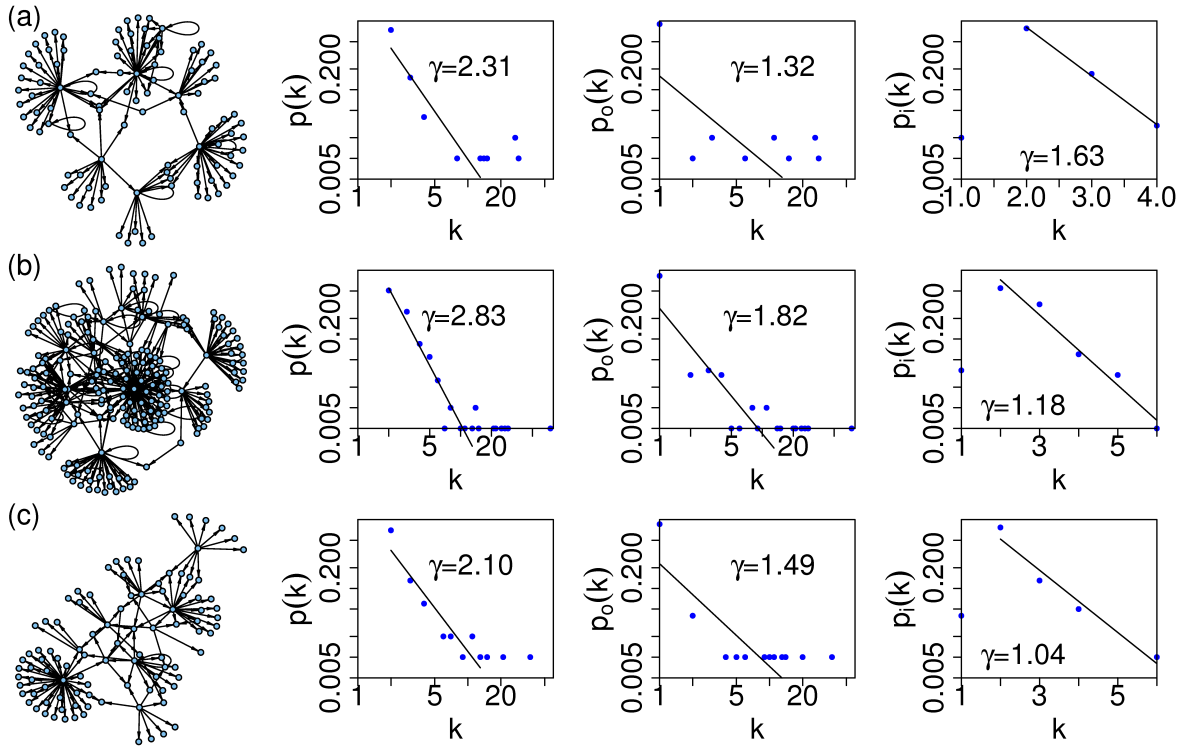

Figure 4.8: (a) Illustration of the network and its degree distribution for 100 genes in *E.* coli. Here and in the following figures $p(k)$ is the frequency of nodes with total degree k, $p_o(k)$ is the frequency of nodes with an out-degree k (both shown in double logarithmic plot), and $p_i(k)$ is the frequency of nodes with an in-degree k (shown logarithmic on the ordinate). Furthermore, the network and its degree distribution for (b) 200 genes of *E.* coli, and (c) 100 genes of *S.* cerevisiae are shown.

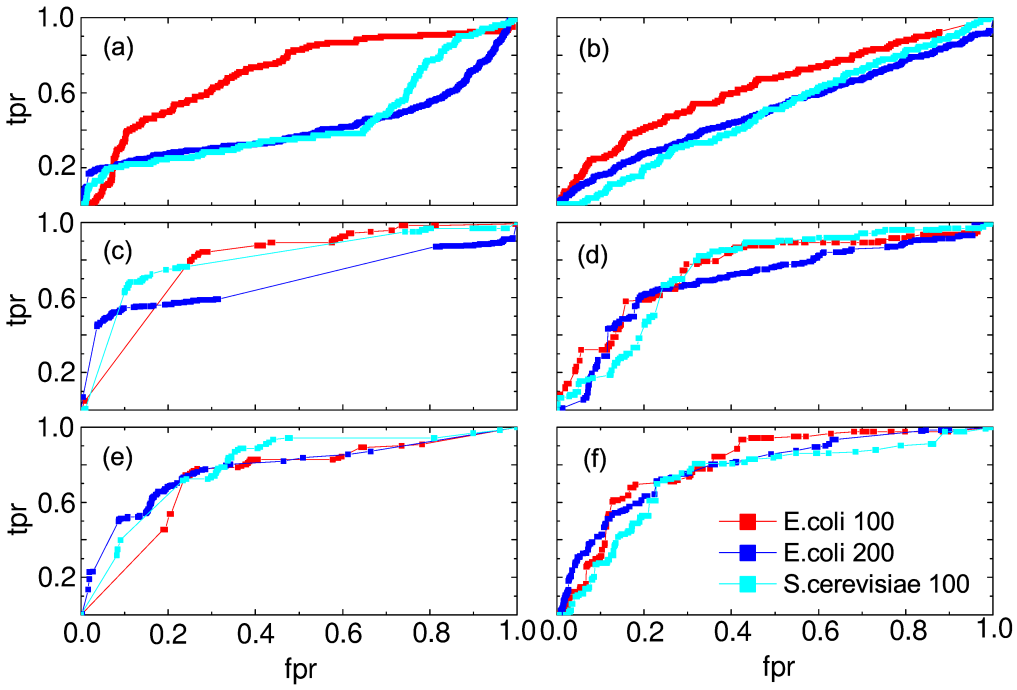

Figure 4.9: ROC curves obtained from the reconstruction of an *E. coli* network of 100 genes, a *S. cerevisiae* network of 100 gene and an *E. coli* network of 200 genes using various association measures: (a) partial Pearson correlation μ_P^p, (b) conditional Granger causality μ_G^c, (c) Spearman correlation μ_S, (d) simple mutual information μ_I, (e) symbol sequence similarity μ_T^S, and (f) residual mutual information μ_I^r.

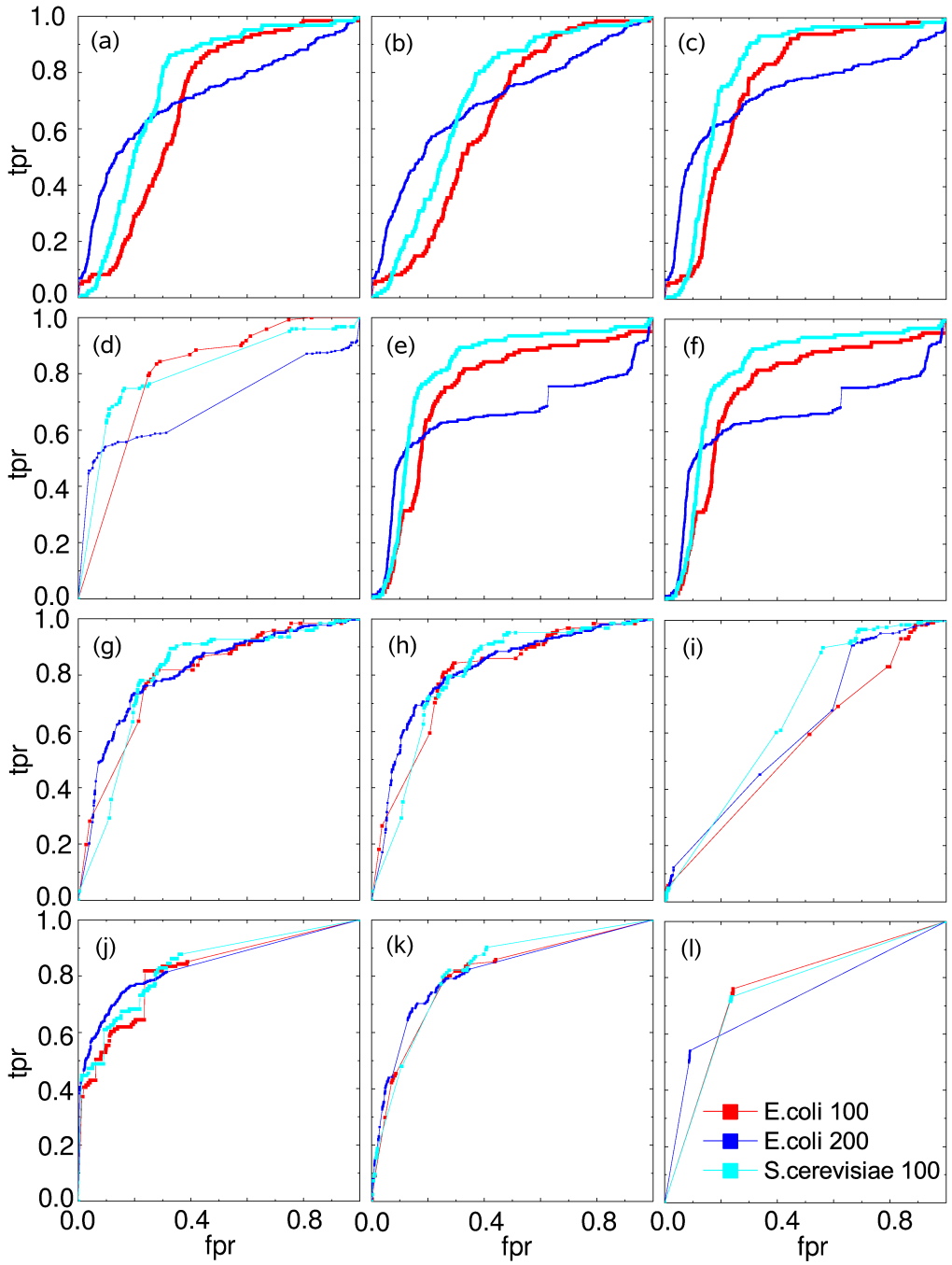

Figure 4.10: ROC curves for various network reconstructions (analog to Fig. 4.9) with the ID scoring scheme and (a) Euclidean distance μ_{EC}, (b) Manhattan distance μ_{MA}, (c) L^s norm μ_L, (d) Kendall's rank correlation μ_K, (e) Pearson's correlation μ_P, (f) conditional Pearson correlation μ_P^c, (g) mutual information of symbol vectors μ_T^I, (h) mean of symbol sequence similarity and the mutual information of symbol vectors μ_T^{SI}, and (i) conditional mutual information μ_I^c. Moreover, the results with Kendall's rank correlation μ_K and (j) $MRNET$, (k) CLR, and (l) $ARACNE$ scoring schemes are shown.

5 IOTA – a novel association measure for reconstructing directed networks

In the previous chapters, I elucidated that most of the currently available association measures, *e.g.*, information-theoretic, correlation, or model-based ones [CC07, GR02, Sch00, HSPVB07, Li90, GSK$^+$08, NRT$^+$10, HKNK11] often may not resolve the network reconstruction problem due to the bias caused by the limited number of time points in molecular biology studies. On the other hand, I showed that measures operating on symbolic dynamics and ranks appear less sensitive to the length of the time series [WSR$^+$09, HKNK11]. Nevertheless, only few measures[1], operating exclusively on long time series, address the important problem of the directionality of coupling [Sch00, DCB07], essential in inferring directed networks. Moreover, none of the measures takes autoregulation into consideration.

In order to overcome these problems and to reduce the number of false positives, I developed the Inner cOmposiTion Alignment ($IOTA$, denoted by ι) [HKKN11], a **novel** asymmetric permutation-based measure, as well as several variants and extensions thereof. This includes an extension in order to determine the type of regulatory links, *i.e.*, activation (positive coupling) or inhibition (negative coupling), particularly important for biological systems. Furthermore, I define a partial variant to discriminate between direct and indirect links in a network and to detect autoregulation [Alo07]. Thus, $IOTA$ has the following merits:

- It can infer statistically significant (nonlinear) couplings from short time series.
- It is capable to infer bi- or unidirectional coupling together with its directionality.
- It allows to infer the type of regulation (activation or inhibition).
- It can distinguish indirect from direct coupling and indicate autoregulation.

This renders $IOTA$ the only existing association measure that can determine **all** necessary characteristics when reconstructing GRN's.

In what follows, I give the definition of the measure and investigate separately the properties of $IOTA$ and variants thereof, particularly with respect to the most common data transformations.

5.1 Inner composition alignment

Inner composition alignment ($IOTA$) is a **novel** permutation-based normalized asymmetric association measure, which I recently proposed to infer directed networks from short time-resolved data [HKKN11, HKNa]. The measure is based on the idea of data reordering in the

[1] Directed (asymmetric) measures are, *e.g.*, Granger causality [GSK$^+$08], coarse-grained entropy rates [PKHS01] or transfer entropy [Sch00].

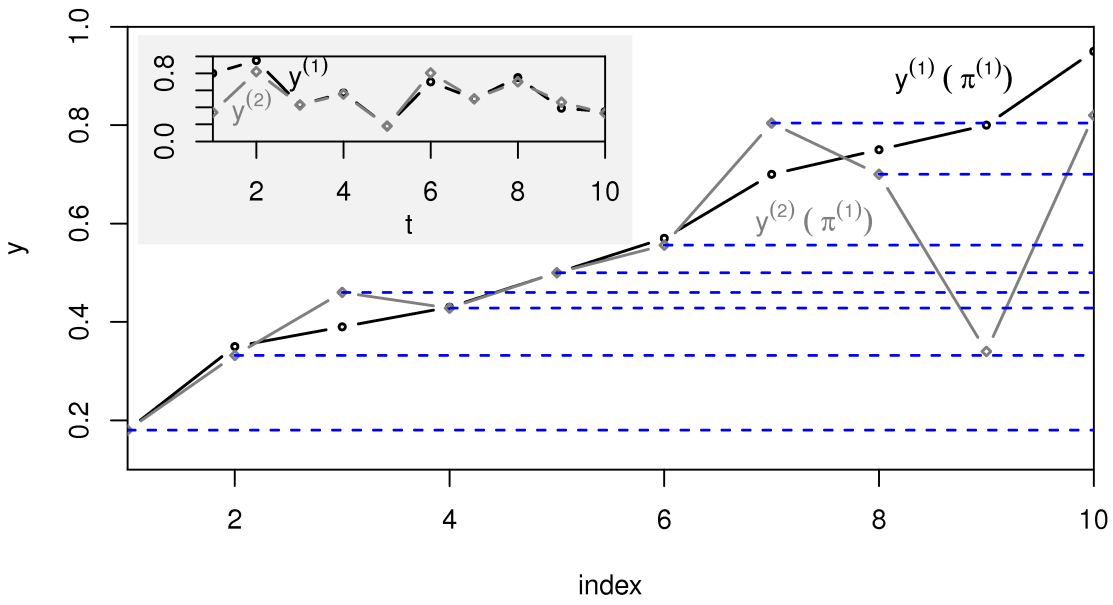

Figure 5.1: The principle of inner composition alignment ($IOTA$): time series $y^{(1)}$ and $y^{(2)}$ are reordered by the permutation $\pi^{(1)}$. Horizontal lines are drawn at points of $y^{(2)}(\pi^{(1)})$. The inset plot shows the time series in their original order.

time series of one subsystem to represent a monotonically increasing function, and then using this particular permutation to reorder the time series of a second subsystem of the network. The conclusion whether both subsystems are linked is then drawn based on the similarity of the reordered time series. In contrast to the rank and symbol based measures studied before, $IOTA$ can obtain different values depending on whether the first or the second series is employed to define the permutation. This renders $IOTA$ an asymmetric measure which facilitates the identification of bi- or unidirectional relationships and enables reconstructing the directionality of the coupling.

5.1.1 Defining the pairwise measure

Given the time series $y^{(l)}$ and $y^{(k)}$ over the same time domains, let $\pi^{(l)}$ be the permutation which orders $y^{(l)}$ in an increasing order, meaning

$$\pi^{(l)} : \ \forall i \ [y^{(l)}(\pi^{(l)})]_i \leq [y^{(l)}(\pi^{(l)})]_{i+1}. \tag{5.1}$$

The series $g^{(k,l)} = y^{(k)}(\pi^{(l)})$ is the reordering of the time series $y^{(k)}$ with respect to $\pi^{(l)}$. The crucial point here is that (for gene expression), for two interacting subsystems, the reordered time series have been observed to be monotonically increasing functions [VdBVLN+06a].

To quantify the monotonicity of the reordered time series, the number of intersection points

name	formula
uniform weighting:	1
arithmetic mean:	$\frac{1}{2}\left(g_{j+1}^{(k,l)} + g_{j}^{(k,l)}\right)$
geometric mean:	$\sqrt{g_{j+1}^{(k,l)} \cdot g_{j}^{(k,l)}}$
harmonic mean:	$2\left(\frac{1}{g_{j+1}^{(k,l)}} + \frac{1}{g_{j}^{(k,l)}}\right)^{-1}$
maximal excursion:	$\max\left(\left\|g_{j+1}^{(k,l)} - g_{i}^{(k,l)}\right\|, \left\|g_{j}^{(k,l)} - g_{i}^{(k,l)}\right\|\right)$
slope:	$\left\|g_{j+1}^{(k,l)} - g_{j}^{(k,l)}\right\|$
squared slope:	$\left(g_{j+1}^{(k,l)} - g_{j}^{(k,l)}\right)^2$

Table 5.1: Different weighting functions w_{ij} for the inner composition alignment

with the horizontal lines which are drawn from each of the time points (Fig. 5.1) are counted. Without loss of generality all lines are drawn dexterwise. Therefore, in order to estimate the probability of the existence of the link $(l \to k)$ the measure μ_ι can be compute by Eq. (5.2):

$$\mu_\iota^{(l\to k)} = 1 - \frac{\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} w_{ij}\Theta[(g_{j+1}^{(k,l)} - g_{i}^{(k,l)})(g_{i}^{(k,l)} - g_{j}^{(k,l)})]}{\Delta}, \tag{5.2}$$

where n is the length of the time series,

$$\Delta = \frac{(n-1)(n-2)}{2} \tag{5.3}$$

is a normalization constant which corresponds to the maximal number of crossings, w_{ij} represents a weight and $\Theta[x]$ is the Heaviside step function

$$\Theta[x] = \begin{cases} 1\ , & x > 0 \\ 0\ , & x \leq 0 \end{cases}\ . \tag{5.4}$$

For two subsystems which are coupled, the number of crossing points tends to zero, rendering a value close to one for μ_ι.

The monotonicity of the reordered time series can be perturbed by external influences and noise, which leads to excursions of the reordered time series. However, these fluctuations are expected to be small compared to those of the reordered time series of independent subsystems. In order to account for noise-induced fluctuations which disturb the monotonicity of the reordered time series and to make the measure more robust, the properties of $IOTA$ are compared using different weighting functions (summarized in Tab. 5.1). The values of the time series have to be rescaled to the interval $[0, 1]$ before calculating $IOTA$ to ensure a proper normalization of the

measure also for non-uniform weighting. Unless otherwise stated, the definition

$$w_{ij} = (g_{j+1}^{(k,l)} - g_j^{(k,l)})^2,\ 0 < i < j < n \tag{5.5}$$

is chosen as weighting function. The reason for that choice will be discussed in Section 7.1.2.

Besides the definition of $IOTA$ given above, it would be also reasonable to draw the lines from each point of the reordered time series $g^{(k,l)}$ to the left. If the $g^{(k,l)}$ are monotonic functions that does not affect the value of $IOTA$. Furthermore, with increasing length of the time series the differences between the values of $IOTA$ obtained with lines drawn to the right and to the left can be expected to decrease fast. Hence, often it is sufficient to count only the intersection points with lines drawn to one side (usually rightwards). However, for decreasing length of the time series larger differences can occur between the values of $IOTA$ in both cases. Hence, for short time series for statistical reasons it is valuable to include both sides. Therefore, $\mu_{\tilde{\iota}}$ is calculated by:

$$\mu_{\tilde{\iota}}^{(l\to k)} = 1 - \frac{\sum_{i=3}^{n}\sum_{j=2}^{i-1} w_{ij}\Theta[(g_{j-1}^{(k,l)} - g_i^{(k,l)})(g_i^{(k,l)} - g_j^{(k,l)})]}{\Delta} \tag{5.6}$$

with lines drawn leftwards, in contrast to the previous definition in Eq. (5.2).

By combining the definitions in Eq. (5.2) and (5.6), the bidirectional inner composition alignment $biIOTA$, μ_{b_ι}, is computed by:

$$\mu_{b_\iota}^{(l\to k)} = \frac{\mu_\iota + \mu_{\tilde{\iota}}}{2}. \tag{5.7}$$

Since this leads to an increase in computational effort, it is not always applicable, in particular if the regulatory relationship between many subsystems must be inferred.

5.1.2 Identifying the type of regulation: inhibition vs. activation

In order to understand the functionality of a regulatory network the direction of coupling and the type of interaction must be investigated. There are two types of regulation to be distinguished: positive or activating and negative or inhibitory coupling interaction. Previously, this kind of information was available only when correlation measures were employed in the network reconstruction process.

$IOTA$ can be modified to address this problem as well:

$$\mu_{\iota e}^{(l\to k)} = T \cdot \mu_\iota^{(l\to k)} \tag{5.8}$$

where

$$T = \operatorname{sign}\left(\frac{\sum_{i=1}^{n-1}(g_{i+1}^{(k,l)} - g_i^{(k,l)})}{n-1}\right). \tag{5.9}$$

The sign of the slopes of the reordered time series $g^{(k,l)}$ is introduced as a factor T in the definition of the measure in order to determine the type of interaction.

The same applies for $\mu_{\tilde{\iota}}$ and μ_{b_ι}. Since per definition the time series that determine the

ordering are always in a nondecreasing order, a positive slope of $g^{(k,l)}$ (*i.e.*, $T > 0$) indicates activation, whereas a negative (*i.e.*, $T < 0$) indicates inhibition.

5.1.3 General properties

In what follows, I exame several mathematical properties of the **novel** measure. As already noted, $IOTA$ is a permutation-based asymmetric association measure, which incorporates the ordering information from one time series and its effect on the second one. Thus, although the time compontent is in general lost in the reordering process, it can indicate the directionality of coupling.

The measure can be defined as a function of two time series $y^{(k)}$ and $y^{(l)}$:

$$\mu_\iota = \mu(y^{(k)}, y^{(l)}), \tag{5.10}$$

where $\mu(y^{(k)}, y^{(l)})$ meets the requirements in Eq. (2.1). Furthermore, for reasons of comparability (as many other association measures) $IOTA$ is per definition normalized to the interval $[0, 1]$ and hence additionally fulfills Eq. (2.2). Due to the introduced nonuniform weighting functions, $IOTA$ requires time series with values in the range between 0 and 1, in order to fulfill Eq. (2.2). If the time series do not meet that requirements, then an appropriate preprocessing in terms of rescaling is needed.

To prove that the $IOTA$ measure meets the requirements of a normalized association measure, expressed by Eq. (2.1) and Eq. (2.2), let the values of the time series $y^{(k)}$ of subsystem k be restricted to the interval $[0, 1]$ for all n time points,

$$[y^{(k)}]_i \in [0, 1] \ , \ \forall i = 1, ..., n. \tag{5.11}$$

Consequently, the values $g_i^{(k,l)}$ (reordered time series) are restricted to the same interval:

$$g_i^{(k,l)} = [y^{(k)}(\pi^{(l)})]_i \in [0, 1] \ , \ \forall i = 1, ..., n. \tag{5.12}$$

Hence, the weights cannot obtain values which are larger than 1:

$$w_{ij} \leq 1 \ , \ \forall i, j, \tag{5.13}$$

and thus the sums in Eq. (5.2) cannot exceed an upper limit of Δ:

$$\begin{aligned} &\sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} w_{ij} \Theta[(g_{j+1}^{(k,l)} - g_i^{(k,l)})(g_i^{(k,l)} - g_j^{(k,l)})] \\ &\leq \sum_{i=1}^{n-2} (n - 1 - i) = \frac{(n-2)(n-1)}{2} = \Delta. \end{aligned} \tag{5.14}$$

Moreover, the sums must be nonnegatve:

$$\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} w_{ij}\Theta[(g_{j+1}^{(k,l)} - g_i^{(k,l)})(g_i^{(k,l)} - g_j^{(k,l)})] \geq 0. \tag{5.15}$$

which follows from Eq. (5.12), since the weights are nonnegative. By inserting the limits obtained in Eq. (5.14) and Eq. (5.15) into the definition of $IOTA$ in Eq. (5.2) it follows that the values of $IOTA$ are restricted to the interval $[0, 1]$. Furthermore, for $y^{(k)} = y^{(l)}$ the reordered time series is defined as $g^{(k,k)} = y^{(k)}(\pi^{(k)})$. That means

$$g_j^{(k,k)} \leq g_{j+1}^{(k,k)} \ , \ \forall j = 1...(n-1), \tag{5.16}$$

$$(g_{j+1}^{(k,k)} - g_i^{(k,k)})(g_i^{(k,k)} - g_j^{(k,k)}) \leq 0 \ , \ \forall j > i \tag{5.17}$$

$$\Theta[(g_{j+1}^{(k,k)} - g_i^{(k,k)})(g_i^{(k,k)} - g_j^{(k,k)})] = 0 \ , \ \forall j > i \tag{5.18}$$

and eventually

$$\mu_\iota^{(k\to k)} = 1. \tag{5.19}$$

Hence, $IOTA$ is a normalized association measure which requires time series with values within the interval $[0, 1]$.

5.1.4 Invariance structure

Data preprocessing (or data transformation) is a common step before analyzing experimentally obtained time series data. For instance, the data may not be normally distributed and may have inhomogeneous variances, invalidating some of the model assumptions used in the analysis (e.g., performing parametric statistical tests such as Student's t-test). Moreover, the usage of statistical tests on time series data which do not meet the model requirements may often lead to misleading results.

In order to overcome these problems, it is necessary to apply data transformation (i.e., a deterministic mathematical function) on the available data set (e.g., z-transform or logarithmic transformations). Therefore, when working with an association measure (here, $IOTA$), it is crucial to examine their invariance with respect to the most common data transformations, namely: translation, scaling, inversion and rotation.

Additionally, as shown before, for the introduced "squared slope" weighting (Eq. (5.5)), it is required that the values of the time series are in range $[0, 1]$ (in order to fulfill Eq. (2.2)). Thus, an appropriate preprocessing in terms of rescaling is necessary. Knowing the invariance structure is essential to avoid misinterpretation of the results.

Translation invariance

$IOTA$ is invariant with respect to a shift of the time series (*e.g.*, by subtracting the average value), both with a uniform weighting and with the one introduced in Eq. (5.5).

Let μ'_ι be a function of the shifted time series $(y^{(k)} + b)$ and $(y^{(l)} + b)$

$$\mu'_\iota = \mu(y^{(k)} + b, y^{(l)} + b) \tag{5.20}$$

where b is constant. Since b does not affect the natural order within the time series, the reordered series are given by

$$g' = (y^{(k)} + b)(\pi^{(l)}) = y^{(k)}(\pi^{(l)}) + b = g + b. \tag{5.21}$$

Hence, the differences between two points of the reordered series are preserved

$$g'_i - g'_j = g_i + b - (g_j + b) = g_i - g_j. \tag{5.22}$$

This in turn renders $IOTA$ to be translation invariant, because both, the argument of the Heaviside step function and the weights, depend only on theses differences.

Scaling invariance

While $IOTA$ is always translation invariant, whether it is invariant with respect to scaling of the time series (*e.g.*, by dividing by the standard deviation) or not depends on the chosen weighting function. Let μ'_ι be a function of the scaled time series $ay^{(k)}$ and $ay^{(l)}$

$$\mu'_\iota = \mu(ay^{(k)}, ay^{(l)}) \tag{5.23}$$

where a is a positive scalar constant. Since scaling with a does not affect the natural order within the time series, the reordered series can be written as

$$g' = (ay^{(k)})(\pi^{(l)}) = ag. \tag{5.24}$$

and the values of the Heaviside step function do not change

$$\begin{aligned} &\Theta[(g'_{j+1} - g'_i)(g'_i - g'_j)] \\ &= \Theta[a^2(g_{j+1} - g_i)(g_i - g_j)] \\ &= \Theta[(g_{j+1} - g_i)(g_i - g_j)]. \end{aligned} \tag{5.25}$$

Hence, in case of a uniform weighting $IOTA$ is invariant with respect to scaling.

However, the weighting function that was introduced to render the measure more robust against noise breaks the invariance, because

$$w'_{ij} = (g'_{j+1} - g'_j)^2 = a^2(g_{j+1} - g_j)^2 = a^2 w_{ij}. \tag{5.26}$$

Thus, scaling all time series with the same factor affects the values of the pairwise weights $\mu_\iota^{(k,l)}$,

but keeps the order of these weights:

$$\begin{aligned}\mu'_\iota &= 1 - \frac{\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} a^2 w_{ij}\Theta[(g_{j+1}-g_i)(g_i-g_j)]}{\Delta} \\ &= a^2\mu_\iota + 1 - a^2. \end{aligned} \tag{5.27}$$

It is important to be aware of the preprocessing, since scaling the time series with different factors can change also the order of the obtained pairwise weights. A proper scaling has to be chosen accoring to the research question.

Inversion invariance

Next, μ'_ι is considered to be a function of the inverted time series $-y^{(k)}$ and $-y^{(l)}$

$$\mu'_\iota = \mu(-y^{(k)}, -y^{(l)}) \tag{5.28}$$

which should not contain ties. In that case, the ordering of the time series is inverted leading to the following relation for the reordered series:

$$[g']_i = -[g]_{n+1-i}, \tag{5.29}$$

where n is the number of time points. Thus, with

$$r = n + 1 - i \tag{5.30}$$

$$s = n + 1 - j \tag{5.31}$$

the summation in the equation for $IOTA$ can be rewritten as

$$\begin{aligned}&\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} w'_{ij}\Theta[(g'_{j+1}-g'_i)(g'_i-g'_j)] \\ &= \sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} w'_{ij}\Theta[(g_{n+2-j}-g_{n+1-i})(g_{n+1-i}-g_{n+1-j})] \\ &= \sum_{r=3}^{n}\sum_{s=2}^{r-1} w_{rs}\Theta[(g_{s-1}-g_r)(g_r-g_s)] \end{aligned} \tag{5.32}$$

where the weights

$$w'_{ij} = (g'_{j+1}-g'_j)^2 = (g_{n+2-j}-g_{n+1-j})^2 = (g_{s+1}-g_s)^2 = w_{rs} \tag{5.33}$$

are unaffected by the inversion. From Eq. (5.32) it is obvious that $IOTA$ is not invariant with respect to inversion, since

$$\mu'_\iota = \mu_{\tilde{\iota}}. \tag{5.34}$$

However, this renders μ_{b_ι} to be invariant with respect to inversion.

Rotation invariance

Finally, let μ'_ι be a function of the rotated time series $ay^{(k)} - b^{(k)}$ and $ay^{(l)} - b^{(l)}$

$$\mu'_\iota = \mu(ay^{(k)} - b^{(k)}, ay^{(l)} - b^{(l)}) \tag{5.35}$$

where a, $b^{(k)}$ and $b^{(l)}$ are constant. From the previous analysis it follows that for $a = 1$

$$\mu'_\iota = \mu_\iota \tag{5.36}$$

while for positive $a \neq 1$ this is only true with uniform weighting. For the squared slope weighting

$$\mu'_\iota = 1 - a^2 + a^2\mu_\iota \tag{5.37}$$

is obtained. Furthermore, for $a = -1$

$$\mu'_\iota = \mu_{\tilde{\iota}} \tag{5.38}$$

applies and for negative $a \neq -1$ in combination with the non-uniform weighting introduced in Eq. (5.5) the relation is

$$\mu'_\iota = 1 - a^2 + a^2\mu_{\tilde{\iota}}. \tag{5.39}$$

Hence, an unrestricted rotation invariance is obtained only for μ_{b_ι} with uniform weighting. If a non-uniform weighting function is used, a proper scaling has to be chosen accoring to the research question. However, for the reconstruction of GRN's relative changes of the concentration of a protein (or a pre-product such as mRNA) are analyzed, where the necessary concentration to observe a specific effect differs among the proteins. Hence, a different scaling of the time series is warrantable in that case.

5.1.5 Statistical properties

The statistical analysis of the proposed association measure is based on permutation tests, as described next. Note that if the same permutation is applied to all time series, the value of μ_ι does not change. Hence, random permutations of all time series are used to discriminate between the following cases:

1. In general, a larger value of the pairwise measure $\mu_{\iota^{(l\to k)}}$ gives a higher probability that two subsystems are linked, leading to the following hypothesis for the permutation test

$$H_0 : E[\mu_{\iota_r}^{(l\to k)}] \geq \mu_\iota^{(l\to k)} \tag{5.40}$$
(no significant coupling),

$$H_A : E[\mu_{\iota_r}^{(l\to k)}] < \mu_\iota^{(l\to k)}. \tag{5.41}$$

Here $\mu_{\iota_r}^{l\to k}$ is the value of $IOTA$ obtained for two randomized time series.

2. If $\mu_\iota^{(l\to k)} > \mu_\iota^{(k\to l)}$, the probability is higher that the subsystem l regulates k. I evaluate $\mu_{\iota_d}^{(l\to k)} = \mu_\iota^{(l\to k)} - \mu_\iota^{(k\to l)}$ and $\mu_{\iota_{d,r}}^{(l\to k)} = \mu_{\iota_r}^{(l\to k)} - \mu_{\iota_r}^{(k\to l)}$ on the supposition of the following hypothesis:

$$H_0 : E[\mu_{\iota_{d,r}}^{(l\to k)}] \geq \mu_{\iota_d}^{(l\to k)} \tag{5.42}$$
(no significant tendency for $l \to k$),

$$H_A : E[\mu_{\iota_{d,r}}^{(l\to k)}] < \mu_{\iota_d}^{(l\to k)}. \tag{5.43}$$

If H_0 is true, then the link is kept as bidirectional, since the driving and the response system cannot be distinguished.

5.2 A partial variant

Next, the definition of $IOTA$ is revisited to identify superfluous links, *i.e.*, to distinguish indirect from direct coupling, and detect possible autoregulatory links. The partial variant of $IOTA$ is realized by applying two consecutive permutations as discussed in the following. Given the subsystem m which regulates the subsystems k and l directly, any pairwise measure, including $IOTA$, will predict an additional link from k to l with a certain probability. To verify whether this link is indirect, the permutations $\pi^{(k)}$ and $\pi^{(m)}$ are determined. Then, it is evaluated whether applying the permutation composition $\pi^{(k)}(\pi^{(m)})$ on $y^{(l)}$ instead of the permutation $\pi^{(m)}$ alone changes the value of the measure. Hence, the partial version of $IOTA$ (Eq. (5.44)) is formulated by comparing the triplets deduced from the pairwise measure:

$$\mu_{\iota^p}^{((k\to l)|(m\to k),(m\to l))} = \left|\mu_\iota(h^{(l,k,m)}) - \mu_\iota(g^{(l,m)})\right|, \tag{5.44}$$

with $g^{(l,k)} = y^{(l)}(\pi^{(k)})$ and $h^{(l,k,m)} = y^{(l)}(\pi^{(k)}(\pi^{(m)}))$. The value of μ_{ι^p} then tends to zero if $(k \to l)$ is an indirect link. For statistical purposes, μ_ι can be replaced with $\mu_{\tilde{\iota}}$ or μ_{b_ι} to calculate the corresponding partial measures.

Similarly, for $k = l = m$ an autoregulatory link can be inferred by

$$\mu_{\iota^p}^{(k\circlearrowleft)} = \left|\mu_\iota(h^{(k,k,k)}) - \mu_\iota(g^{(k,k)})\right|, \tag{5.45}$$

with $g^{(k,k)} = y^{(k)}(\pi^{(k)})$ and $h^{(k,k,k)} = y^{(k)}(\pi^{(k)}(\pi^{(k)}))$. In this case, a low value of μ_{ι^p} is obtained if the time series are almost monotonic, which on the other hand indicates a low probability for autoregulation. Again, μ_ι can be replaced by $\mu_{\tilde{\iota}}$ or μ_{b_ι}. However, the differentiation between positive and negative autoregulation is not possible.

5.2.1 Invariance structure of the partial measure

Next, I investigate the properties of the partial measure with respect to the four data transformations, which were discussed previously for the pairwise variants of $IOTA$. For this, let μ_{ι^p}

be a function of the time series $y^{(k)}$, $y^{(l)}$ and $y^{(m)}$

$$\mu_{\iota^p} = \mu(y^{(k)}, y^{(l)}, y^{(m)}). \tag{5.46}$$

Translation invariance

In order to examine whether μ_{ι^p} is translation invariant, consider μ'_{ι^p} as a function of the shifted time series $(y^{(k)} + b)$, $(y^{(l)} + b)$ and $(y^{(m)} + b)$:

$$\mu'_{\iota^p} = \mu(y^{(k)} + b, y^{(l)} + b, y^{(m)} + b), \tag{5.47}$$

where b is constant. Similar to Eq. (5.21), b does not affect the natural order within the time series. Thus, the reordered time series are given by:

$$g' = (y^{(l)} + b)(\pi^{(m)}) = y^{(l)}(\pi^{(m)}) + b = g + b \tag{5.48}$$

$$h' = (y^{(k)} + b)(\pi^{(l)}(\pi^{(m)})) = y^{(k)}(\pi^{(l)}(\pi^{(m)})) + b = h + b \tag{5.49}$$

Combining the previous equations and Eq. (5.44), it follows that the partial version of $IOTA$ is invariant with respect to translation of the corresponding time series.

Scaling invariance

Now, let μ'_{ι^p} be a function of the scaled time series $(ay^{(k)})$, $(ay^{(l)})$ and $(ay^{(m)})$

$$\mu'_{\iota^p} = \mu(ay^{(k)}, ay^{(l)}, ay^{(m)}) \tag{5.50}$$

where a is a positive scalar constant. Because the natural order within the time series is unaffect by the scaling, the reordered series are given by

$$g' = (ay^{(l)})(\pi^{(m)}) = ag \tag{5.51}$$

$$h' = (ay^{(k)})(\pi^{(l)}(\pi^{(m)})) = ah \tag{5.52}$$

If the squared slope weighting is applied, it follows from Eq. (5.27) that

$$\mu'_{\iota^p} = \left| a^2\iota(h) + 1 - a^2 - a^2\iota(g) - 1 + a^2 \right| = a^2\mu_{\iota^p}. \tag{5.53}$$

For uniform weighting μ_{ι^p} is invariant with respect to the scaling.

Inversion invariance

In the case of inversion of the investigated time series, the relations for the reordered series are as follows

$$[g']_i = -[g]_{n+1-i} \tag{5.54}$$

$$[h']_i = -[h]_i. \tag{5.55}$$

This implies that the partial version of $IOTA$ is not invariant with respect to inversion of the time series. The same is also true for $biIOTA$ (Eq. 5.7).

Rotation invariance

Due to the effect of scaling and inversion, μ_{ι^p} is not invariant with respect to rotation.

5.2.2 Statistical properties of the partial measure

The significance of the regulatory links is quantified by a permutation test with random permutations of all time series. In order to remove superfluous links within the reconstructed network, the partial variant of $IOTA$ is applied to the triplets $(k \to l), (m \to k), (m \to l)$, where $\mu_{\iota^p}^{((k\to l)|(m\to k),(m\to l))}$ is expected to tend to zero if the link from l to k is indirect. Therefore, the following hypotheses need to be proven:

1. If the link $k \to l$ is not necessary to explain the regulation of l, then $\mu_\iota(h^{(l,k,m)})$ and $\mu_\iota(g^{(l,m)})$ do not differ significantly:

$$H_0 : E[|\mu_{\iota_r^p}^{((k\to l)|(m\to k),(m\to l))}|] \geq |\mu_{\iota^p}^{((k\to l)|(m\to k),(m\to l))}| \tag{5.56}$$

(l and k are not significantly dependent, that means the link $(k \to l)$ is superfluous),

$$H_A : E[|\mu_{\iota_r^p}^{((k\to l)|(m\to k),(m\to l))}|] < |\mu_{\iota^p}^{((k\to l)|(m\to k),(m\to l))}|. \tag{5.57}$$

Here $\mu_{\iota_r^p}$ is the partial measure calculated from the randomized time series.

2. Additionally, autoregulation can be considered if $k = l = m$, given the hypothesis:

$$H_0 : E[|\mu_{\iota_r^p}^{(k\circlearrowleft)}|] \geq |\mu_{\iota^p}^{(k\circlearrowleft)}| \tag{5.58}$$

(indicating almost monotonic time series, that means no significant autoregulation),

$$H_A : E[|\mu_{\iota_r^p}^{(k\circlearrowleft)}|] < |\mu_{\iota^p}^{(k\circlearrowleft)}|. \tag{5.59}$$

Hence, the partial measures can identify time series of autoregulated systems.

5.3 Comparison to Kendall's rank correlation

Concepts of time series reordering are common and widely used to study dependencies between (sub)systems. In this context, rank correlations have been shown to reliably infer coupling from short time series (Chapter 2) and to be robust with respect to noise (Chapter 4). Therefore, in the following, the similarities and differences between $IOTA$ (in particular, μ_ι) and Kendall's

τ are elucidated, where the latter is based on a similar idea as $IOTA$ and has been shown to infer reliably undirected networks (Chapter 2). For these purpose, the definition of Kendall's τ is reformulate in the following manner:

Given two time series, the Kendall's rank correlation can be calculated by determining the permutations, which arrange the respective series in nondecreasing order independently from one another, namely $\pi^{(1)}$ for $y^{(1)}$ and $\pi^{(2)}$ for $y^{(2)}$. If the corresponding values in $\pi^{(1)}$ and $\pi^{(2)}$ are linked (as illustrated in Fig. 5.2 upper panel), then the number of intersections among these links matches the number of discordant pairs:

$$n_d = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \Theta[(r_j^{(1)} - r_i^{(1)})(r_i^{(2)} - r_j^{(2)})] = \frac{n(n-1)}{2} - n_c, \tag{5.60}$$

where $r^{(l)}$ is the rank of time series $y^{(l)}$ related with the permutation $\pi^{(l)}$ and n_c is the number of concordant pairs:

$$n_c = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \Theta[(r_i^{(1)} - r_j^{(1)})(r_i^{(2)} - r_j^{(2)})] \tag{5.61}$$

If down-regulation is assumed, the number of intersections will coincide with the maximum number ($c_{max} = \frac{n(n-1)}{2}$), whereas for up-regulation it will be zero ($c_{min} = 0$). By shifting this range of possible values to the interval $[-1, 1]$, changing the sign by

$$\frac{\frac{c_{max}}{2} - c}{\frac{c_{max}}{2}} = 1 - \frac{2c}{c_{max}} = 1 - \frac{4c}{n(n-1)} \tag{5.62}$$

and setting $n_d = c$, the definition of Kendall's τ can be written in the following form:

$$\mu_K = 2\frac{n_c - n_d}{n(n-1)} = 1 - \frac{2}{n(n-1)} \cdot 2n_d. \tag{5.63}$$

The inner composition alignment is based on a similar idea as the Kendall's rank correlation, namely the reordering of time series, however, it is not a correlation measure. Unlike Kendall's measure, $IOTA$ includes the ordering information from one subsystem and its effect on the second one. Again, the permutations $\pi^{(l)}$ are determined. Applying $\pi^{(1)}$ on the ranks of $y^{(1)}$ and $y^{(2)}$ leads to the series $\rho^{(1)} = r^{(1)}(\pi^{(1)})$ and $\rho^{(2)} = r^{(2)}(\pi^{(1)})$ which are subsequently used to calculate μ_ι:

$$\mu_\iota^{(1\to2)} = 1 - \frac{2}{n(n-1)} \cdot \frac{n \cdot \kappa}{(n-2)}, \tag{5.64}$$

where the number of crossings in Fig. 5.1 equals

$$\kappa = \sum_{k=1}^{n-2} \sum_{i=k+1}^{n-1} \sum_{j=i+1}^{n} \Theta[(\rho_k^{(2)} - \rho_j^{(2)})(\rho_i^{(2)} - \rho_k^{(2)})] \cdot \Theta[\rho_i^{(1)} - \rho_j^{(1)} + 2] \cdot w_{ij}. \tag{5.65}$$

Thus, the graphical representation of Kendall's τ and $IOTA$ are comparable (*i.e.*, $\rho = r$) only when $y^{(1)}$ are monotonically increasing time series.

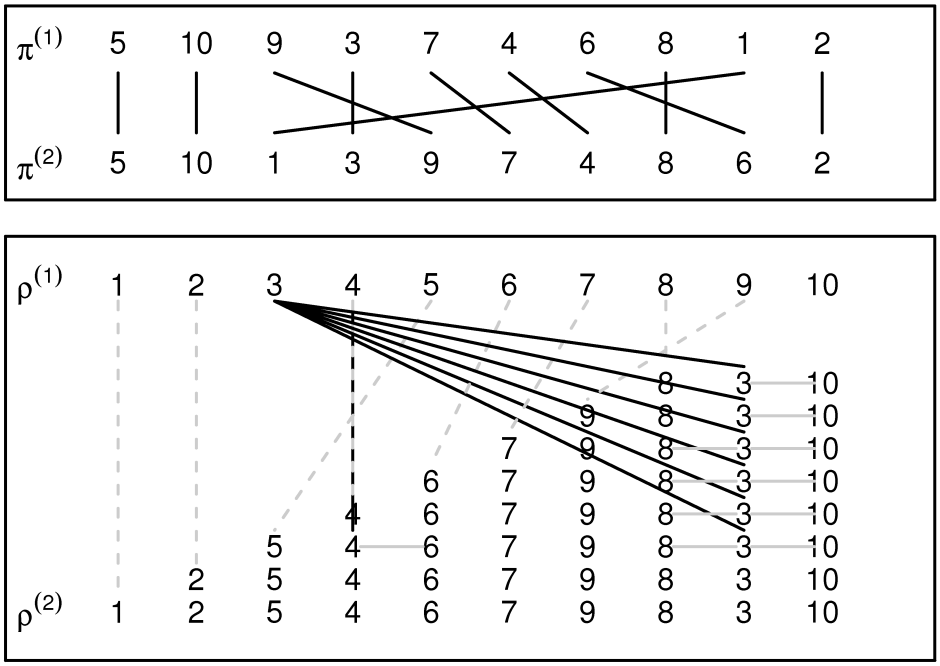

Figure 5.2: Kendall: In the upper panel $\pi^{(1)}$ and $\pi^{(2)}$ are the permutations sorting time series $y^{(1)}$ and $y^{(2)}$ (displayed in Fig. 5.1) in increasing order. The number of intersection points corresponds to the discordant pairs ($n_d = 8$). *IOTA*: In the lower panel $\rho^{(1)}$ and $\rho^{(2)}$ are shown, which are the ranks of $y^{(1)}$ and $y^{(2)}$ both reordered according to the permutation $\pi^{(1)}$. All values in $\rho^{(2)}$ are linked to the associated in $\rho^{(1)}$. For reasons of clarity the values of $\rho^{(2)}$ are reprinted dexterwise from the link in a separate row for each link. Next, for each fixed value $\rho_i^{(2)}$ pairs of neighbors in $\rho^{(2)}$ (both to the right of $\rho_i^{(2)}$) are linked, if there associated values in $\rho^{(1)}$ are at opposite sites of $\rho_i^{(1)}$ (associated to $\rho_i^{(2)}$). The amount of horizontal lines corresponds to the number of crossings in Fig. 5.1 ($\kappa = 11$). Thus, μ_K (Eq. 5.63) and μ_ι (Eq. 5.64) are different.

6 A general numerical study on IOTA's capabilities for coupling analysis

In order to evaluate the capabilities of $IOTA$ for inferring directed networks, next, I apply the measure on various simulated time-resolved data sets including autoregressive processes and chaotic oscillators. Moreover, the influence of the length and the type of time series[1], as well as the dependence on the coupling type (*i.e.*, homogenous or heterogenous in the network, uni- or bidirectional, activating or inhibitory) is investigated.

6.1 Application to paradigmatic network modules

First, to examine the properties of the novel measure in detail and to evaluate its capabilities to reconstruct directed networks, I apply $IOTA$ to several small paradigmatic network modules, which serve as toy models.

6.1.1 Case study 1

A crucial factor which influences the accuracy of the data-driven reconstruction of complex networks is the length of the available time series. In general, the accuracy of the association measures decreases for shorter time series. Therefore, various measures involve different biases when they are applied to very short time series. On the other hand, it has been shown in Chapter 2 that rank and symbol based measures are less affected by the length of the time series [HKNK11]. Hence, $IOTA$ is expected to be robust with respect to the length of the time series and it can be applied to very short data sets.

In order to examine the actual dependence on the length of the time series, initially I study a small network module of 3 subsystems for very short (10 time points) and longer time series (> 3000 time points). The system is defined by:

$$y_i^{(1)} = -0.3y_{i-1}^{(1)} + u_i^{(1)} \tag{6.1}$$

$$y_i^{(2)} = \left(y_{i-1}^{(1)}\right)^2 - 0.5y_{i-1}^{(2)} \tag{6.2}$$

$$y_i^{(3)} = u_i^{(3)} \tag{6.3}$$

where the first subsystem is driving the second, and both of them are autoregulated (here, negatively regulating themselves to keep the values of the time series bounded). The first two subsystems together represent unidirectionally coupled $AR(1)$ processes while the third is

[1] In that context, the term "short time series" has to be broaden depending on the type of the time series.

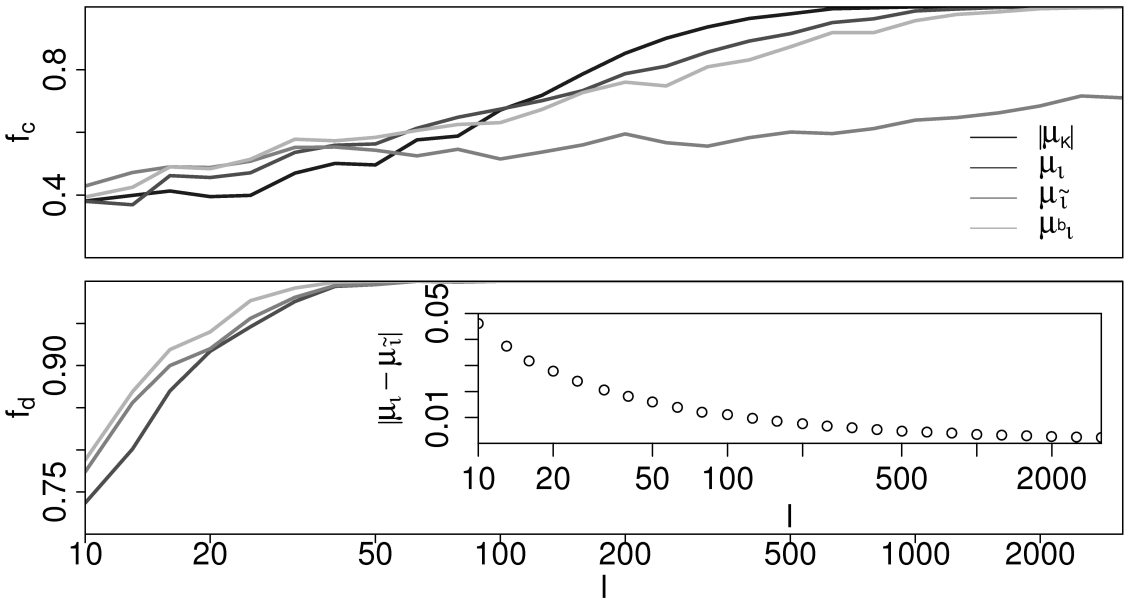

Figure 6.1: Fraction f_c of realizations where the coupling is identified correctly (upper panel), and fraction f_d where the correct coupling direction is inferred (lower panel). The results are shown for the pairwise $IOTA$ measure and the different versions thereof, as well as for Kendall's rank correlation measure. Note that for Kendall's τ f_d is zero per definition. The inset plot in the lower panel shows how the difference between μ_ι ($IOTA$ with lines rightwards) and $\mu_{\tilde{\iota}}$ ($IOTA$ with lines leftwards) evolves with varying length of the time series.

independent from the previous ones. Various time series are generated: for each length 10^a (a varied from 1.0 to 3.5 in steps of 0.1) 1000 realizations of the time series are considered, where $u_i^{(k)} = N(0,1)$ are standard, normally distributed random values.

Next, the different pairwise variants of $IOTA$ as well as Kendall's rank correlation are applied to reconstruct the links in this system. Moreover, the values of the pairwise measures are stored in a matrix $I = \mu(y^{(k)}, y^{(l)})$, in order to compare the efficiency of the two $IOTA$ variants (*i.e.*, when its values are obtained with lines drawn rightwards and leftwards) and of $biIOTA$ to that of Kendall's τ. Here μ is either μ_ι, $\mu_{\tilde{\iota}}$, μ_{b_ι} or $|\mu_K|$. Since these measures do not address autoregulation, the diagonal of the matrix I is set to zero in all cases.

Hence, a measure is considered to perform well in identifying the coupling if its value is maximal for the true link, that is, if $\mu(y^{(1)}, y^{(2)}) = \max(I)$ in case of the model system in Eq. (6.1)–(6.3). Furthermore, a measure indicates the correct coupling direction ($k \to l$), if $\mu(y^{(k)}, y^{(l)}) > \mu(y^{(l)}, y^{(k)})$, regardless of its value with respect to the maximum $\max(I)$ over all pairs of time series.

For all investigated lengths of time series the fraction of realizations where the coupling is identified correctly (f_c) and the fraction where the correct coupling direction is inferred (f_d) are

evaluated. Here it has to be noted that for Kendall's rank correlation ($|\mu_K|$), the fraction f_d is zero per definition, since the measure is symmetric.

Figure 6.1 illustrates that for short time series, the capabilities of all measures to correctly identify the coupling are very similar, where $\mu_{\bar{\iota}}$ ($IOTA$ with lines drawn leftwards) and μ_{b_ι} ($biIOTA$) perform slightly better than μ_ι and Kendall's τ. In addition, in contrast to the symmetric correlation measure, $IOTA$ is capable to indicate also the direction of coupling. Even for very short time series, the correct coupling direction is identified in more than 70% of the cases. For intermediate length of the time series (100–1000 time points) Kendall's τ performs best in identifying the coupling closely followed by μ_ι ($IOTA$ with lines drawn rightwards) and μ_{b_ι}. However, the different variants of $IOTA$ additionally indicates the correct coupling direction. Finally, for time series of order $> 10^3$, μ_ι and μ_{b_ι} can identify the correct coupling for all time series realizations, whereas for $\mu_{\bar{\iota}}$ the fraction of correctly inferred links converges much slower towards the maximal value of 1.

This difference in the convergence of μ_ι and $\mu_{\bar{\iota}}$ reflects the observation that for random time series, the number of crossings from lines drawn leftwards tends to be slightly larger than from lines drawn rightwards, and hence, $\mu_{\bar{\iota}}$ is smaller than μ_ι. However, the difference between the values obtained for μ_ι and $\mu_{\bar{\iota}}$ decreases significantly as the length of the time series increases (Fig. 6.1 inset plot).

Moreover, the capability of μ_ι, $\mu_{\bar{\iota}}$ and μ_{b_ι} to distinguish the direction of coupling converges rather fast, and for time series of more than 50 time points the correct direction is always identified for the model system used here. Additionally, for short time series, μ_{b_ι} performs best in discriminating the coupling direction, followed by $\mu_{\bar{\iota}}$ and μ_ι. However, all three variants of $IOTA$ can infer the correct coupling direction from short time series in approximately $75-80\%$ of the cases.

6.1.2 Case study 2

To further explore the capabilities of $IOTA$ in reconstructing directed networks from short time series, next, I increase the number of nodes and examine three network modules, where each model system corresponds to a network of 7 nodes representing interacting subsystems and the coupling between the nodes differs among the models (Fig. 6.2). The dynamics of the nodes is discrete, where the value of the time series of one node at time step i is governed by the values of the time series of all its input nodes at time step $i-\xi$.

The following study relies on short time series composed of 10 time points for each node, which is a typical length for biological data from high-throughput experiments as discussed previously in this work. Since all three pairwise variants of $IOTA$ performed similar on such short data sets (as shown in Subsection 6.1.1), only μ_ι is considered here. However, the partial measure is employed additionally to investigate the capabilities of $IOTA$ to identify superfluous links and autoregulation. Moreover, the significance of the links is assessed by the permutation test (empirical p-values at significance level 0.01) which is described in the previous chapter in Subsections 5.1.5 and 5.2.2. In addition, the resulting reconstruction efficiency is compared to that of correlation measures, namely Pearson's, Spearman's and Kendall's correlation coefficient, where I focus in particular on Kendall's rank correlation.

The investigated systems (network modules) are described in Tab. 6.1, where the time point

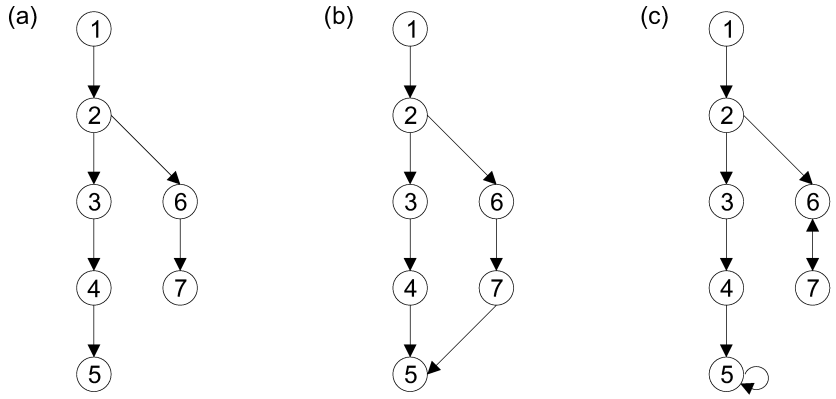

Figure 6.2: Illustration of the 3 network modules corresponding to the 3 models in Tab. 6.1: (a) model 1, (b) model 2, (c) model 3

$\{y_i^{(k)}\}$ of the time series of node k is determined by the expression noted in the column for the particular model. Furthermore, to examine the influence of different coupling situations, time delays and the general dynamical behavior, on the reconstruction efficiency, the models are studied for various parameters.

For model 1 and 2 the same coupling strength $\bar{\epsilon}$ (scaling factor) and time delay $\bar{\xi} = [0, 2] \in \mathbb{Z}$ between all coupled subsystems are used. Moreover, the same exponent q is employed in all equations. Scaling and exponent are varied in the simulations ($\bar{\epsilon} \in [-10^5, -10^{-5}] \cup [10^{-5}, 10^5]$, $q \in [-5, -1] \cup [1, 5]$). Furthermore, the initial time series is modeled with standard, normally distributed random values $u = N(0, 1)$, fixed for all q and $\bar{\epsilon}$.

In model 3, I study the influence of inhomogeneous coupling and delay and compare the results for all $\epsilon_d = 1$ and all time delays $\xi_d = \bar{\xi} \in [0, 2]$ ($\forall d \in [1, 8]$). The exponent q is fixed at $q = -1$ in the simulations. Moreover, in addition to the homogenous delays, several time delays are considered between the coupled nodes, where the following values are included: $\{\xi_{2d} = 2, \xi_{2d-1} = 1\}$ and viseversa, as well as $\{\xi_d = 1, \xi_{d+4} = 2\}$ ($\forall d \in [1, 4]$). Furthermore, the following inhomogeneous coupling situations are examined ($\forall d \in [1, 4]$): $\{\epsilon_d = -1, \epsilon_{d+4} = 1\}$, $\{\epsilon_d = E_d, \epsilon_{d+4} = E_d\}$ and $\{\epsilon_d = -E_d, \epsilon_{d+4} = E_d\}$, where E_d is additionally varied including $E_d = \{1, 2, 3, 4\}$ and all possible permutation thereof.

As illustrated in Fig. 6.3, the empirical study for model 1 indicates that the correlation-based measures fail to detect the direct links for time delay $\bar{\xi} = 1$, while indirect coupling introduces artifact links in the reconstructed network. On the other hand, *IOTA* is capable to detect the direct links for various combinations of the parameter $\bar{\epsilon}$ and q, however, further on it detects the indirect links inferred from the correlation-based measures as well. Additionally, for dense networks, as the one studied here, the directions are often indistinguishable. The results for model 2 confirm the general observations, as shown in Fig. 6.4.

The capabilities of using *IOTA* to correctly infer coupling, however, depends on both the

	model 1	model 2	model 3
$y_i^{(1)}$	u_i	u_i	u_i
$y_i^{(2)}$	$\bar{\epsilon}(y_{i-\bar{\xi}}^{(1)})^q$	$\bar{\epsilon}(y_{i-\bar{\xi}}^{(1)})^q$	$\epsilon_5(y_{i-\xi_5}^{(1)})^q$
$y_i^{(3)}$	$\bar{\epsilon}(y_{i-\bar{\xi}}^{(2)})^q$	$\bar{\epsilon}(y_{i-\bar{\xi}}^{(2)})^q$	$\epsilon_6(y_{i-\xi_6}^{(2)})^q$
$y_i^{(4)}$	$\bar{\epsilon}(y_{i-\bar{\xi}}^{(3)})^q$	$\bar{\epsilon}(y_{i-\bar{\xi}}^{(3)})^q$	$\epsilon_7(y_{i-\xi_7}^{(3)})^q$
$y_i^{(5)}$	$\bar{\epsilon}(y_{i-\bar{\xi}}^{(4)})^q$	$\bar{\epsilon}((y_{i-\bar{\xi}}^{(4)})^q + (y_{i-\bar{\xi}}^{(7)})^q)$	$\epsilon_3(y_{i-\xi_3}^{(4)})^q + \epsilon_1(y_{i-\xi_1}^{(5)})^q$
$y_i^{(6)}$	$\bar{\epsilon}(y_{i-\bar{\xi}}^{(2)})^q$	$\bar{\epsilon}(y_{i-\bar{\xi}}^{(2)})^q$	$\epsilon_2(y_{i-\xi_2}^{(2)})^q + \epsilon_4(y_{i-\xi_4}^{(7)})^q$
$y_i^{(7)}$	$\bar{\epsilon}(y_{i-\bar{\xi}}^{(6)})^q$	$\bar{\epsilon}(y_{i-\bar{\xi}}^{(6)})^q$	$\epsilon_8(y_{i-\xi_8}^{(6)})^q$

Table 6.1: The investigated time series $\{y_i^{(k)}\}$ of node k for the 3 paradigmatic network modules shown in Fig. 6.2.

density of the underlying network and the time delay in the time series. While in general the time information is lost by reordering, the definition of $IOTA$ allows to conserve the temporal relation between the time series. The empirical study reveals that the measure performs particularly well for small time delays (Fig. 6.5), in contrast to correlation-based measures which perform well only for time delay $\bar{\xi} = 0$. Moreover, the presence of small delays increases the capability of $IOTA$ to remove indirect links and to infer the directionality of coupling in dense networks. However, autoregulation can remain undiscovered for small, and rather dense networks.

Additionally, Fig. 6.6 elucidates that the resulting reconstruction efficiency for unidirectional coupling is barely affected by introducing inhomogeneous coupling. Furthermore, even though $IOTA$ can not always determine the correct coupling direction from very short time series, it distinguishes well uni- from bidirectional coupling: while in case of the unidirectionally coupled subsystems $IOTA$ usually keeps only one direction, it correctly detects the bidirectional coupling for several parameter combinations.

6.2 Coupled oscillators

So far, general properties of $IOTA$ have been studied on the basis of toy models with discrete dynamics. However, the question remains if the obtained findings are robust when exchanging the system (more realistic time series). In nature systems of interest (*e.g.*, weather, stockmarkets, laser, or chemical reactions, to just name a few) are often well described only by systems of nonlinear differential equations and, within a certain parameter range, chaotic behavior is frequently observed. Explaining such behavior may be done via the analysis of a mathematical mode. However, usually this requires analytical techniques to infer first the model from time series. Thus, chaotic time series, as important realistic data sets, are interesting applications to further test the capabilities of $IOTA$ for inferring coupling structures.

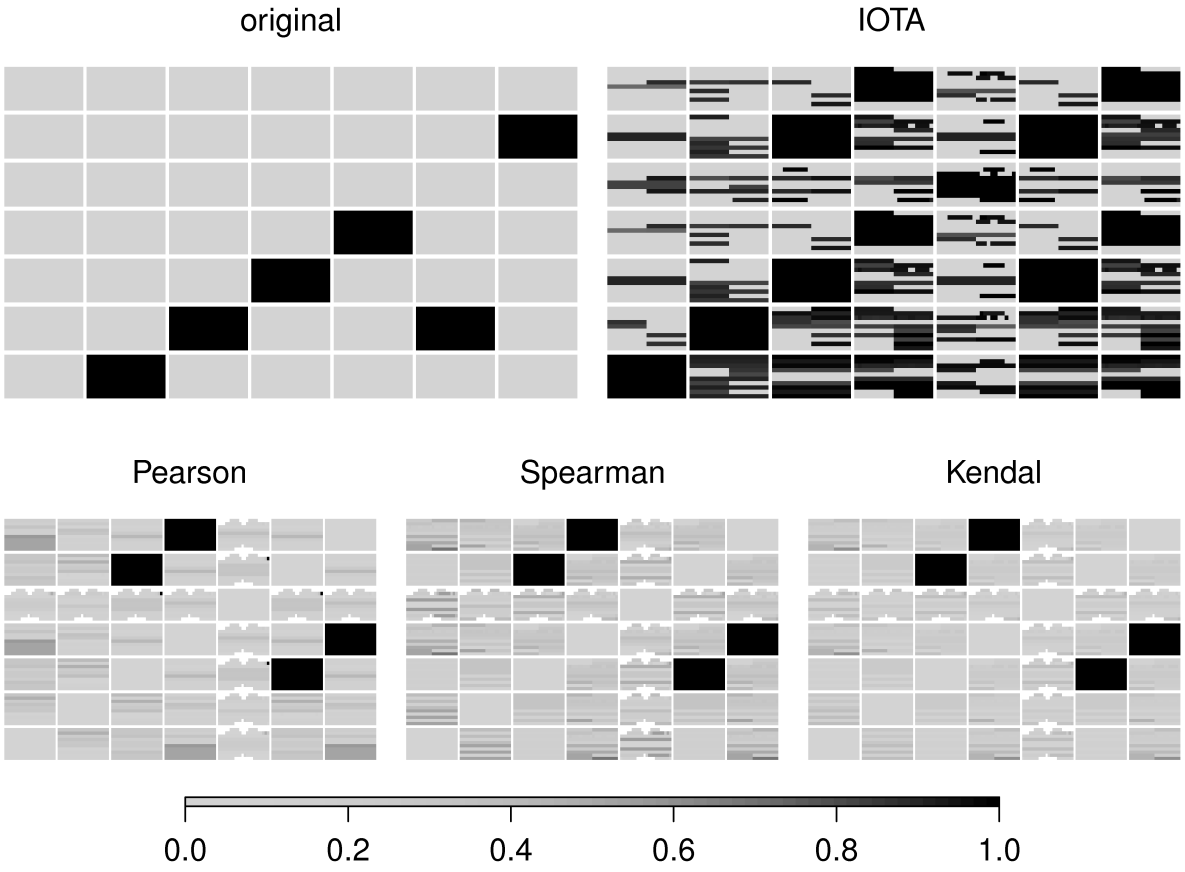

Figure 6.3: Coupling (black high, light gray low) deduced for model 1 (Tab. 6.1) with delay $\bar{\xi} = 1$; each rectangle represents an entry of the adjacency matrix, where the colors within concern the measure's values for various model parameters (q vertical, $\bar{\epsilon}$ horizontal), η is randomly chosen. For Pearson, Spearman and Kendall the absolute value of the correlation coefficient is shown. *IOTA* relies on the lines drawn dexterwise and includes the application of the pairwise and the partial measures.

6.2.1 A network module with chaotic dynamics

First, the influence of the type of the dynamics of the system under investigation on the reconstruction efficiency is investigated. Hence, the important problem of inferring the coupling of chaotic oscillators is reinterpreted to analyze once more a small network module similar to those in the previous section. In particular, the following coupled Roessler-Lorenz system is chosen here to govern the dynamics of a module of six nodes with inhomogeneous coupling:

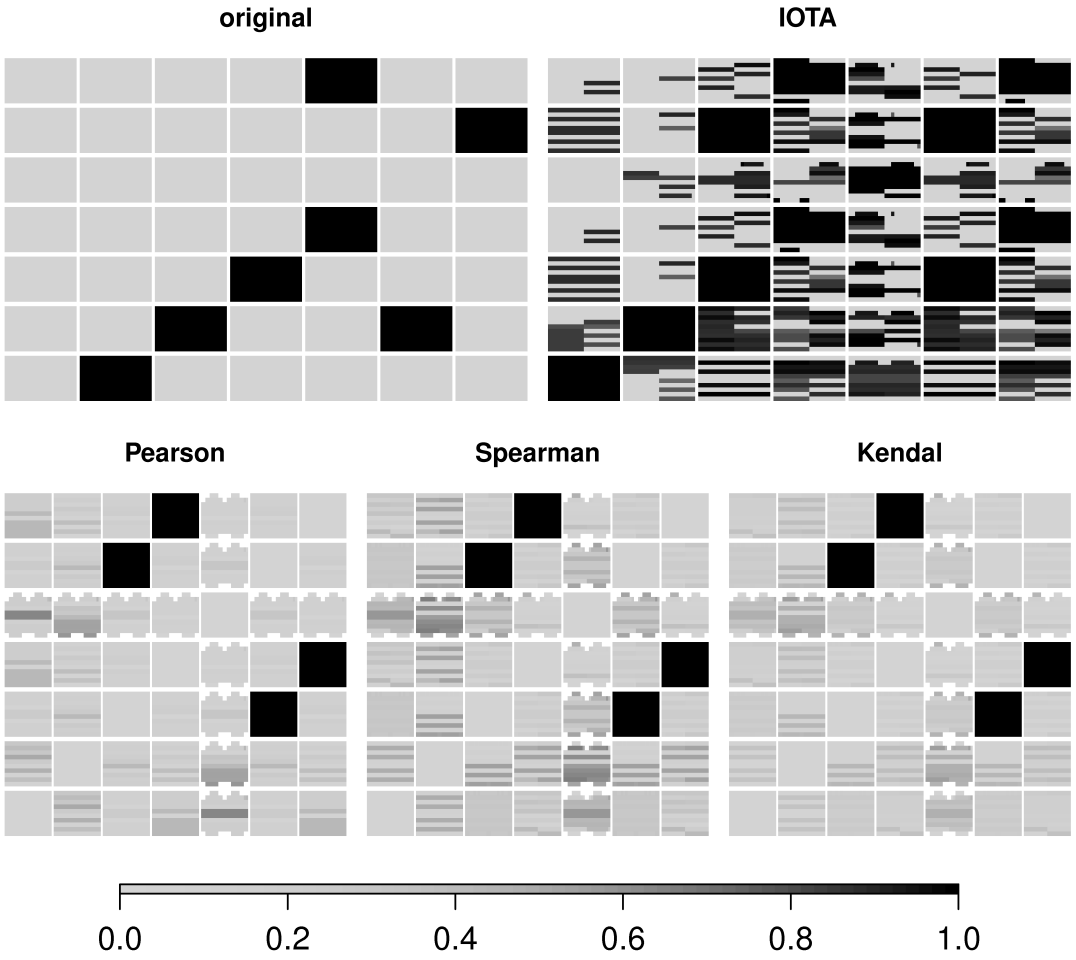

Figure 6.4: Coupling deduced for model 2 (Tab. 6.1) with delay $\bar{\xi} = 1$; illustration analog to Fig. 6.3

$$\dot{y}^{(1)} = -6(y^{(2)} + y^{(3)}) \tag{6.4}$$
$$\dot{y}^{(2)} = 6(y^{(1)} + 0.2y^{(2)}) \tag{6.5}$$
$$\dot{y}^{(3)} = 6(0.2 + y^{(3)}(y^{(1)} - 5.7)) \tag{6.6}$$
$$\dot{y}^{(4)} = 10(y^{(5)} - y^{(4)}) \tag{6.7}$$
$$\dot{y}^{(5)} = 28y^{(4)} - y^{(4)}y^{(6)} - y^{(5)} + \epsilon y^{(2)} \tag{6.8}$$
$$\dot{y}^{(6)} = y^{(4)}y^{(5)} - \frac{8}{3}y^{(6)} \tag{6.9}$$

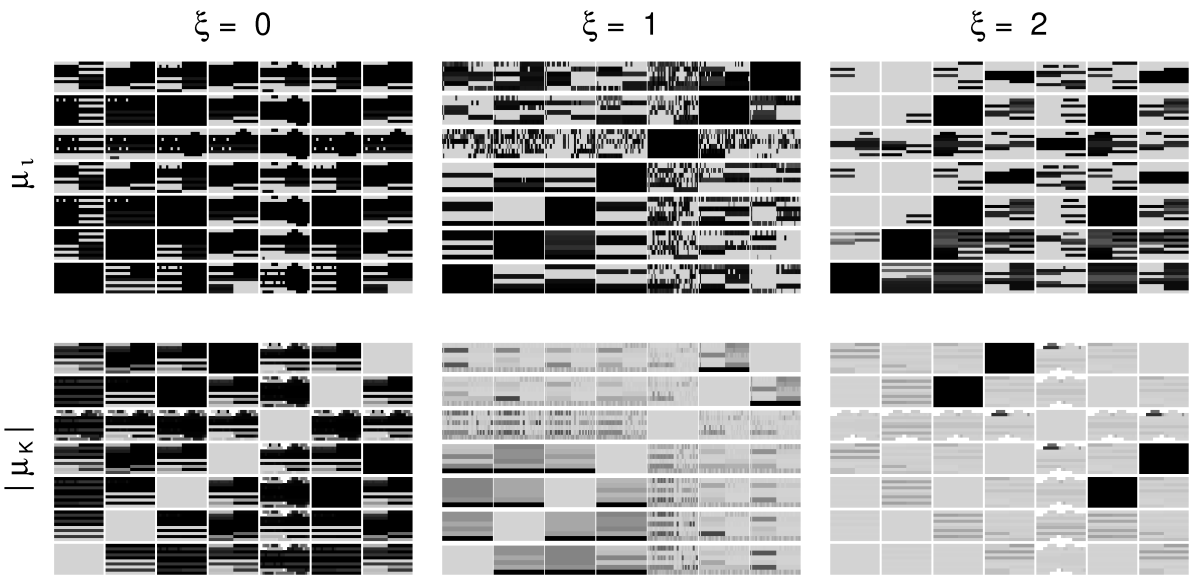

Figure 6.5: Coupling deduced for model 1 (Tab. 6.1) with 3 different values of the delay ξ which is the same between all nodes ($\bar{\xi} = \xi$). The upper panels show the results obtained by *IOTA*, the lower ones those by Kendall's τ; illustration and color coding analog to Fig. 6.3.

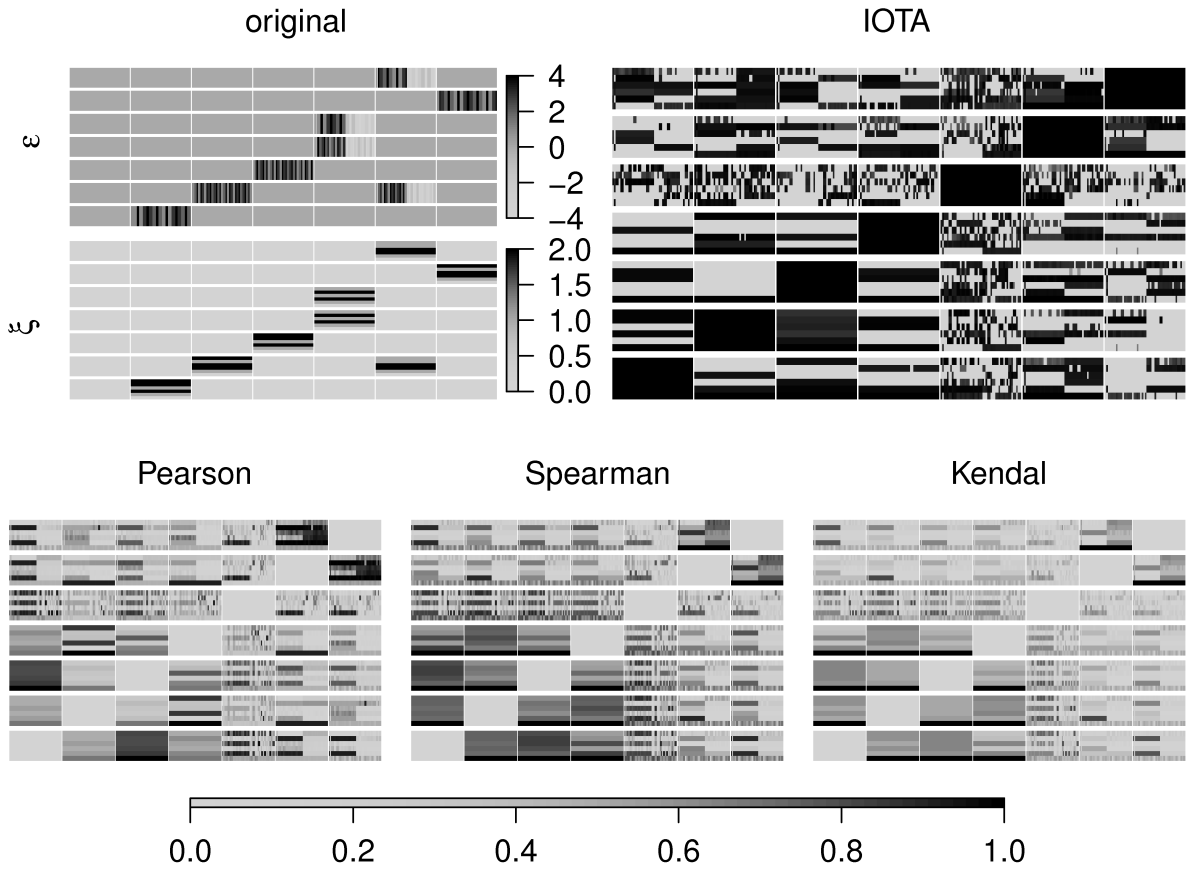

Figure 6.6: In the upper left panel (named original) the colors within each rectangle indicate the parameter values (coupling strength ϵ and time delay ξ) which has been used to compute time series for model 3 (Tab. 6.1). The other panels show the coupling (values of $IOTA$, as well as Pearson's, Spearman's and Kendall's correlation coefficient) deduced for the model; illustration analog to Fig. 6.3.

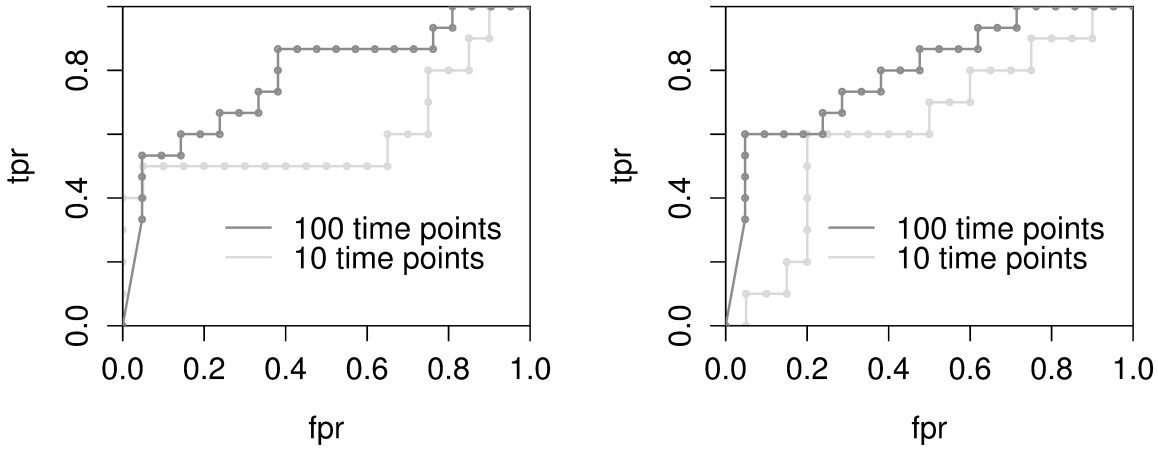

Figure 6.7: ROC curves for the reconstruction of a network module whose dynamics is governed by a coupled Roessler-Lorenz system, with coupling $\epsilon = 10^{-0.64}$ (left) and $\epsilon = 10^{-1.5}$ (right).

with $y^{(1)}, y^{(2)}, y^{(3)}$ corresponding to the Roessler system and $y^{(4)}, y^{(5)}, y^{(6)}$ to the Lorenz system. The system is numerically integrated for $\epsilon = 10^{-0.64}$ (from 1 to 700, sampled at 10 Hz) using a Runge-Kutta scheme of order three with adaptive step size [BS89] and the first 2000 points are neglected. This results in time series of 5000 time points (capturing approximately 11 oscillations).

Short time series consisting of 100 (and respectively 10) time points are extracted by resampling, *i.e.*, choosing every 50-th (500-th) point of the original time series, and $IOTA$ is applied to all pairs of the six time series. Only the pairwise measure μ_ι is computed here.

The obtained ROC curves, shown in Fig. 6.7 (left), suggest that $IOTA$ is also valuable for the inference of chaotic time series. Thus, the general findings obtained in the previous study of toy models translate well for more realistic time series.

For instance, if the mean of the values of μ_ι obtained among all pairs of different time series

$$\bar{\mu}_\iota = \frac{1}{30} \sum_{k \neq l} \mu_\iota(y^{(k)}, y^{(l)}) \tag{6.10}$$

is chosen as a threshold, a tpr of approximately 87% and a fpr of approximately 38% are obtained for the time series of length 100. Performing the same analysis with the pairwise $IOTA$ for time series of length 10 the tpr is approximately 67%, where the fpr is the same as before, but the coupling between the Roessler and the Lorenz system ($y^{(2)} \to y^{(5)}$) is not detected.

Additional simulations with different coupling strengths for the link $y^{(2)} \to y^{(5)}$ revealed that these results are quite sensitive to the choice of the systems parameters. For example, in a second

simulation with weaker coupling $\epsilon = 10^{-1.5}$, the obtained rates of true and false positives are very similar for the time series of lengths 100 and 10 (tpr ≈ 0.73, fpr ≈ 0.38), if the threshold is defined as before. This means the tpr is deteriorated in case of the 100 time points and improved in case of the 10 time points. However, the coupling between the Lorenz and the Roessler system ($y^{(2)} \to y^{(5)}$) is still not detected from the 10 time points, but from the 100. Additionally, the ROC curves in Fig. 6.7 (right) indicate a significantly better performance for the time series of length 100.

All together this implies that *IOTA* is capable to investigate couplings between subsystems with complex dynamics, but it might be also valuable to infer coupling among chaotic systems from short time series. However, the framework conditions needed to obtain proper results have to be further evaluated for various systems and parameters.

6.2.2 Further possible applications - a parameter study of coupled chaotic Roessler oscillators

Next, *IOTA* is applied to investigate the (direct and indirect) interactions in network modules of coupled chaotic oscillators where only the time series of one component per oscillator is available for the analysis [HKN$^+$b], representing a typical situation for many natural systems. The case study in the previous subsection implies that, if the time series are "similar" to each other, *i.e.*, the oscillators are synchronized or close to synchronization, then, the reordered time series of both subsystems can be expected to be monotonically increasing functions. Thus, the values of *IOTA* will be significantly larger than for randomized time series (as described in Section 5.1.5).

In the following, it is explored and discussed how the capability of *IOTA* to infer coupling is affected by the coupling strength, the general coupling situation (uni- and bidirectional coupling) and the choice of system parameters (leading to different dynamical regimes).

For this purpose Roessler oscillators

$$\dot{x}_1^{(k)} = -\omega_k x_2^{(k)} - x_3^{(k)} + \sum_{l \neq k} D_{kl}(x_1^{(l)} - x_1^{(k)}) \tag{6.11}$$

$$\dot{x}_2^{(k)} = \omega_k x_1^{(k)} + a x_2^{(k)} \tag{6.12}$$

$$\dot{x}_3^{(k)} = b + (x_1^{(k)} - c) x_3^{(k)} \tag{6.13}$$

which are coupled diffusively in the first component via the coupling matrix D_{kl} are considered. The parameter $b = 0.1$ and $c = 10$ are chosen universally, while I consider two values for the parameter a which lead to spiral chaos ($a = 0.1625$) in the first case and Funnel chaos ($a = 0.2925$) in the second case. The coupling analysis is performed for a chain of 3 bidirectionally coupled oscillators (Fig. 6.8 (a)) and the results are compared to previous findings obtained with available measures of synchronization.

Additionally, to further investigate the influence of the general coupling situation on the reconstruction efficiency, in the first scenario the bidirectional links are partially replaced by unidirectional ones (Fig. 6.9). Moreover, a star of 4 bidirectionally coupled oscillators (Fig. 6.8 (b)) is investigated to evaluate the influence of the network size.

In all cases, the frequencies ω_k are chosen such that the oscillators are non-identical, namely:

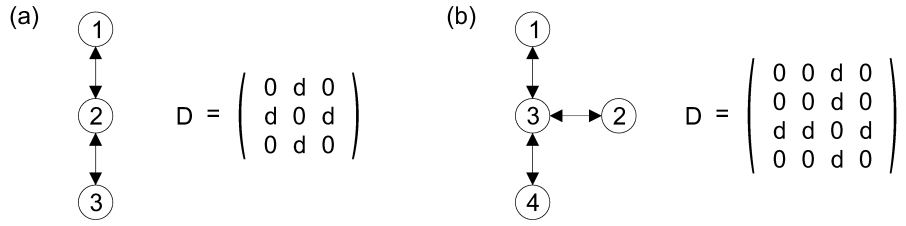

Figure 6.8: Scheme of network modules of bidirectionally coupled Roessler oscillators: (a) three oscillators in a chain, and (b) four oscillators in a star conformation. Additionally the coupling matrix D is shown in both cases.

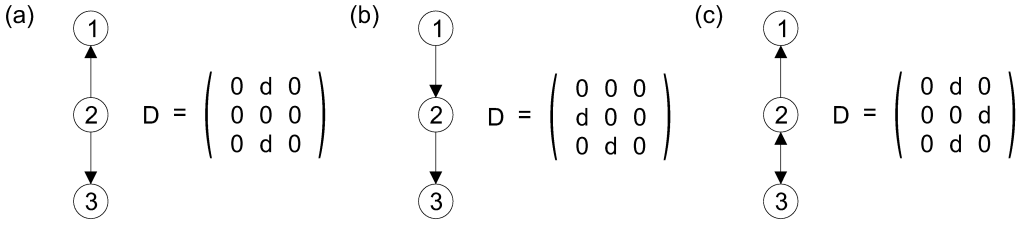

Figure 6.9: Scheme of network modules of (partially) unidirectionally coupled Roessler oscillators in a chain together with the corresponding coupling matrices D: (a) common driver, (b) cascade, (c) mixture of uni- and bidirectional coupling.

$\omega_1 = 0.98$, $\omega_2 = 1.02$, and $\omega_3 = 1.06$ for the chain, and $\omega_1 = 0.98$, $\omega_2 = 1.04$, $\omega_3 = 1.00$, and $\omega_4 = 0.94$ for the star assembly. Moreover, to increase the significance of the results longer time series (1500 time points) are used, in accordance with experimental realizations.

The system is numerically integrated, starting from random initial conditions, by applying a Runge-Kutta scheme of order three with adaptive step size [BS89] until time series of 152000 time points (sampled at 20 Hz) were obtained. Only the time series of the first component of each subsystem are employed for the further analysis ($y^{(k)} = x_1^{(k)}$). To get rid of transient behavior the first 2000 time points are neglected. The remaining parts of the trajectories are splitted into 100 pieces $y^{(k_i)}$ of length 1500, each capturing approximately 11 oscillations.

From these 100 trajectories $IOTA$ is calculated for all pairs of the oscillators and for the various coupling strengths. The obtained values are visualized as boxplots (*e.g.*, in Fig. 6.10) where the medians of $IOTA$ over the 100 time series are represented as horizontal bars and the means are represented as circles (in most cases both are almost overlapping). Furthermore, the boxes correspond to the quartiles (25% to 75% quantile) and the dashed lines show the full range of the obtained values.

The significance of the results is evaluated with a permutation test (Section 5.1.5) where each time series is randomized 100 times. The mean value $\overline{\mu_{\iota_r}^{(k_i \to l_i)}}$ and the variance $\sigma^2\left(\mu_{\iota_r}^{(k_i \to l_i)}\right)$ of $IOTA$ are estimated over these randomizations. Furthermore, I determine the average values

$$\frac{1}{100}\sum_{i=1}^{100}\overline{\mu_{\iota_r}^{(k_i \to l_i)}}$$

and

$$\frac{1}{100}\sum_{i=1}^{100}\sigma\left(\mu_{\iota_r}^{(k_i \to l_i)}\right)$$

over the 100 time series for each coupling strength in order to obtain the 2σ interval (shaded area) which corresponds approximately to a significance level 0.05.

Moreover, $IOTA's$ capability to infer the coupling situation is linked to the state of the coupled system (occurrence of synchronization). To this end, Lyapunov spectra are estimated from the long time series of 152000 time points with TISEAN [HKT99] for pairs of oscillators, since for two coupled oscillators the spectrum and its changes during the synchronization process are well discussed in literature [RPK96].

The Lyapunov spectra, which are estimated with TISEAN, are rough approximations of the full spectrum (because they are estimated from the time series of two oscillators, while the other dimensions of the fully coupled system are neglected). Hence, they are only used to qualitatively distinguishing the states of the coupled systems. The pairwise spectra can be used as indicators to characterize the qualitative differences in the synchronization of the oscillators for different coupling schemes and in the two dynamical regimes.

Bidirectional coupling

First, I consider bidirectional coupling, as the more frequently studied case in literature. In the following investigation, I examine $IOTA's$ capabilities to correctly infer the bidirectional links

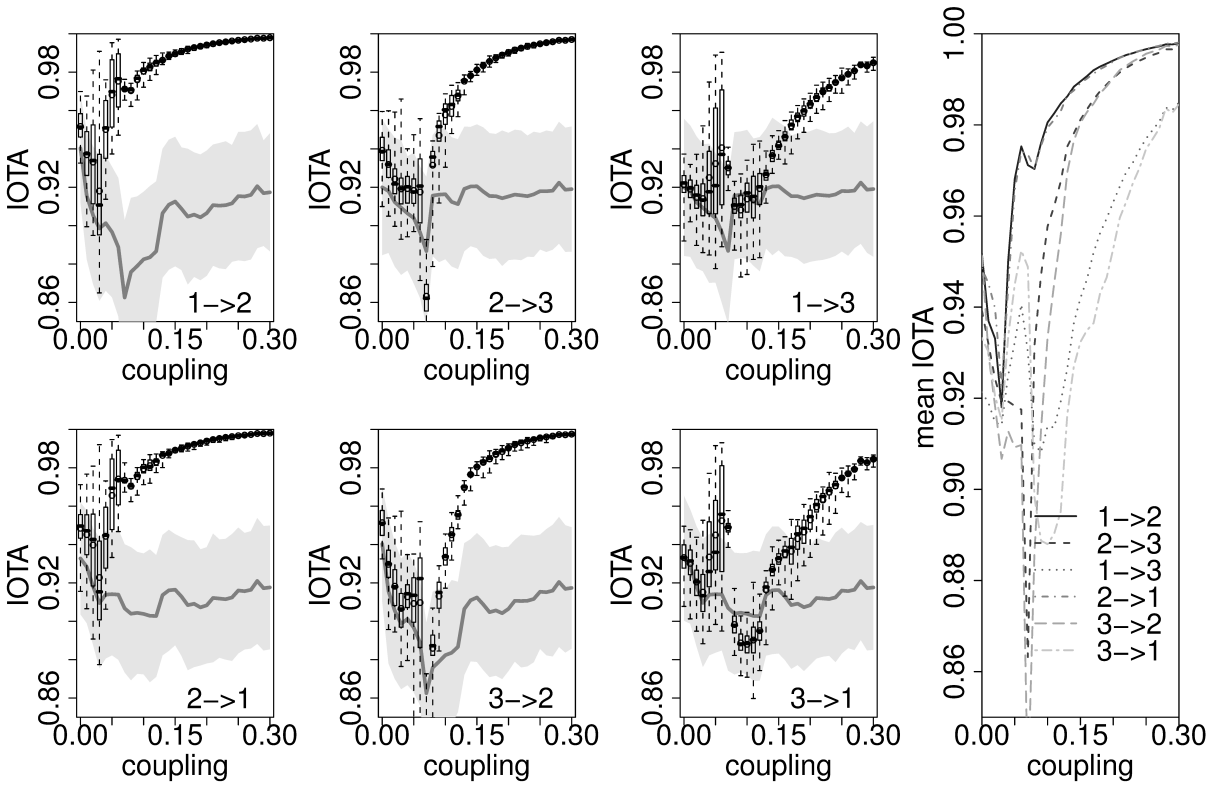

Figure 6.10: *IOTA* for Roessler oscillators (phase-coherent regime) coupled in a chain conformation with bidirectional coupling. Black bars (circles) correspond to the median (mean) of *IOTA* over 100 short trajectories. Boxes mark the quartile, dashed lines the full range of obtained values. Gray illustrates the region of not significant values with respect to the permutation test (2σ interval). In the right panel, the mean of *IOTA* is compared among all pairs.

when the oscillators are in phase-coherent (spiral chaos, $a = 0.1625$) and in non-phase-coherent (funnel chaos, $a = 0.2925$) regime. The results are shown and discussed here for the scenario in Fig. 6.8 (a).

The three Roessler oscillators are arranged in a chain where subsystem 1 and 2, as well as subsystem 2 and 3 are coupled bidirectionally. The coupling d in coupling matrix D (Fig. 6.8 (a)) is varied from 0.0 to 0.3 in steps of 0.01 for the simulation.

For spiral chaos the boxplots (Fig. 6.10) indicate that the link between oscillator 1 and 2 becomes significant already for very low coupling strengths ($d \approx 0.05$), while for the pair 2 and 3 *IOTA* obtains significant values only for a coupling strength larger $d \approx 0.10$. Moreover, the indirect coupling between oscillator 1 and 3 shows significant values basically for $d > 0.14$. These coupling strengths, for which the values of *IOTA* become finally significant, are in accordance with the onset of phase synchronization.

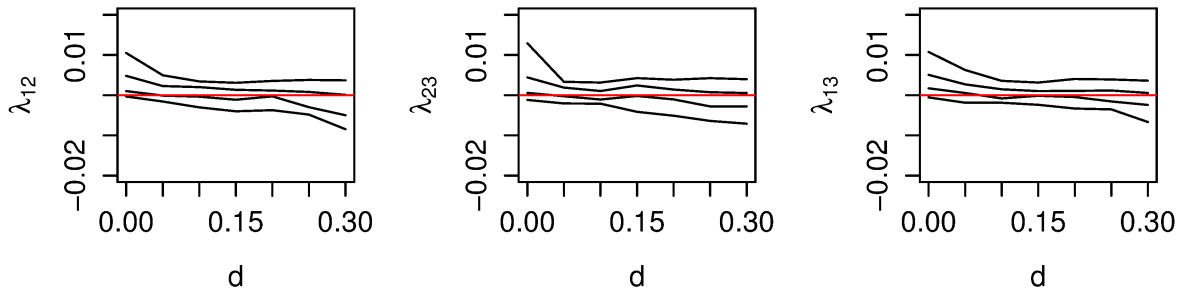

Figure 6.11: Lyapunov spectra for pairs of oscillators in the phase-coherent regime. Approximation of the full Lyapunov spectrum of the chain of 3 Roessler oscillators coupled bidirectionally in a chain conformation. Estimated from the time series of two oscillators, while the other dimensions of the fully coupled system are neglected.

In the estimated Lyapunov spectra[2] for the pairs of oscillators (Fig. 6.11) in all three cases there are two Lyapunov exponents close to zero. The first of them becomes negative almost instantaneously if the coupling strength is larger than zero. The second one becomes negative for coupling strength larger $d \approx 0.15$.

The observed trends in the Lyapunov spectra are in accordance with the conclusion that can be drawn for example from the Hilbert phase[3]. They confirm that in order to obtain significant values of $IOTA$ the subsystems need to be close to phase synchronization (necessary condition). The latter seems to be the fulfilled if the two Lyapunov exponents which are close to zero in the pairwise estimated spectra become negative.

Moreover, for all oscillator pairs, the obtained values of IOTA for both directions (*e.g.*, $1 \to 2$ and $2 \to 1$) are very similar. Additionally, the variations of $IOTA's$ values in the randomized series are similar for both directions which is usually not observed in the case of unidirectional links. Hence, although there are some weak tendencies to prefer one direction, bidirectional coupling is more likely than an unidirectional one.

Furthermore, for $d > 0.14$, where significant coupling is obtained for all oscillator pairs, the values for the indirect links are much lower than for the direct ones (comparing the means, Fig. 6.10 right panel). This indicates that the link between 1 and 3 is the indirect one. The same relation holds true for the partial measure (Fig. 6.12 left panel).

If the oscillators are not in the spiral, but in the Funnel regime much higher coupling strengths are needed to infer the links. However, the boxplots (Fig. 6.13) indicate that the order in which the links become significant is the same as in the phase-coherent case. First, the coupling

[2]The exact onset of phase synchronization cannot be determined, since the estimated Lyapunov exponents are only rough approximations.

[3]The evolution of the differences of the phases Φ_k indicates the onset of phase synchronization at $\bar{d} \approx 0.04$ ($\Phi_1 - \Phi_2$), and $\bar{d} \approx 0.07$ ($\Phi_2 - \Phi_3$ and $\Phi_1 - \Phi_3$) respectively, where the Hilbert phase is employed. In general, for a time series x it is estimated from the analytic signal $z = x + \mathrm{i}y$ as $\Phi_H = \arctan(\frac{y}{x})$, where the imaginary part of the analytic signal is determined by the Hilbert transformation $y(t) = \frac{P}{\pi} \int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau}$ with the Cauchy principal value P. The transformation is performed using the function "hilbert" in the "EMD" package for R.

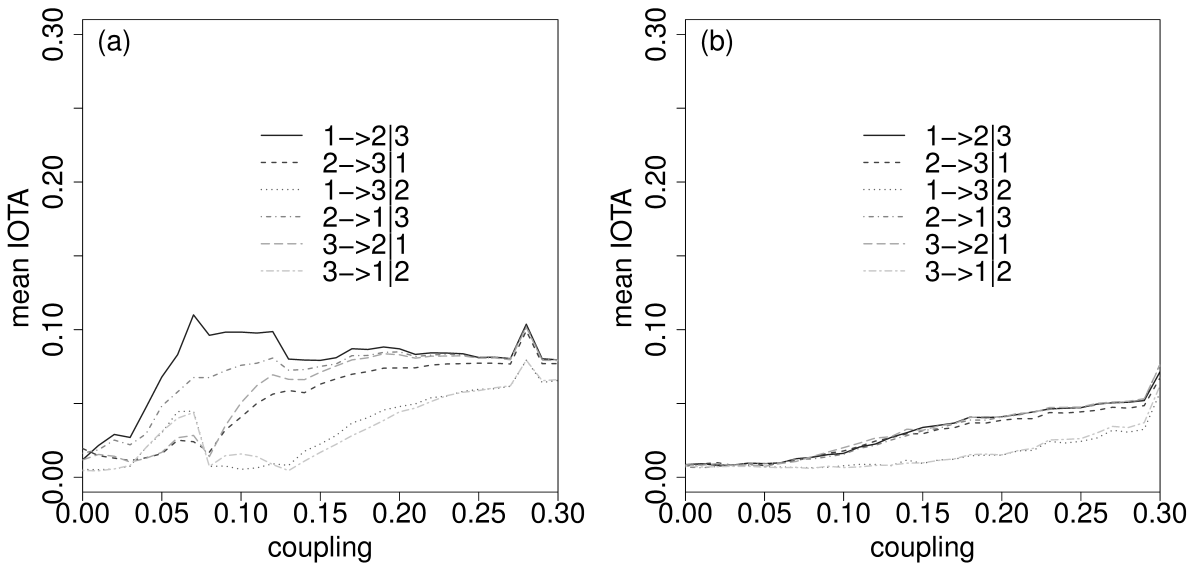

Figure 6.12: Mean of partial $IOTA$ over 100 trajectories. Estimated for bidirectionally coupled Roessler oscillators in a chain conformation, in phase-coherent (left) and non-phase-coherent regime (right).

between oscillator 1 and 2 (for coupling strength $d \approx 0.22$), then the one between 2 and 3 becomes significant (for coupling strength $d \approx 0.25$). Finally, the indirect coupling between oscillator 1 and 3 tends towards significant values. However, for that pair of oscillators no significant values are obtained within the considered range of coupling strengths, although there is a strong tendency that the values of $IOTA$ increase while the coupling strength is increased ($IOTA$ becomes significant with respect to the quartile here).

As already observed in the phase-coherent case, the d values, for which significant coupling is indicated by $IOTA$, are in accordance with the occurrence of synchronization, which is estimated from the Lyapunov spectra in Fig. 6.14 to first occur at coupling strengths of $d \approx 0.25$.

Moreover, for all pairs the directions are barely distinguishable from the absolute values of $IOTA$. Additionally, the variations of $IOTA's$ values in the randomized series are similar for both directions. Hence, also in non-phase-coherent regime the possibility for bidirectional coupling cannot be precluded from the results.

Furthermore, for sufficiently high coupling strength (*i.e.*, significant values of $IOTA$) the direct links obtain much higher values of the measure than the indirect ones, suggesting that the latter can be excluded. The same trend is observed for the partial measure (Fig. 6.12 right panel). However, the distinction between the direct and the indicted links becomes harder for further increased coupling strengths.

Nevertheless, $IOTA$ has been proven a valuable tool for inferring the coupling between 3 (bidirectionally) coupled chaotic subsystems when only short time-resolved measurements are available.

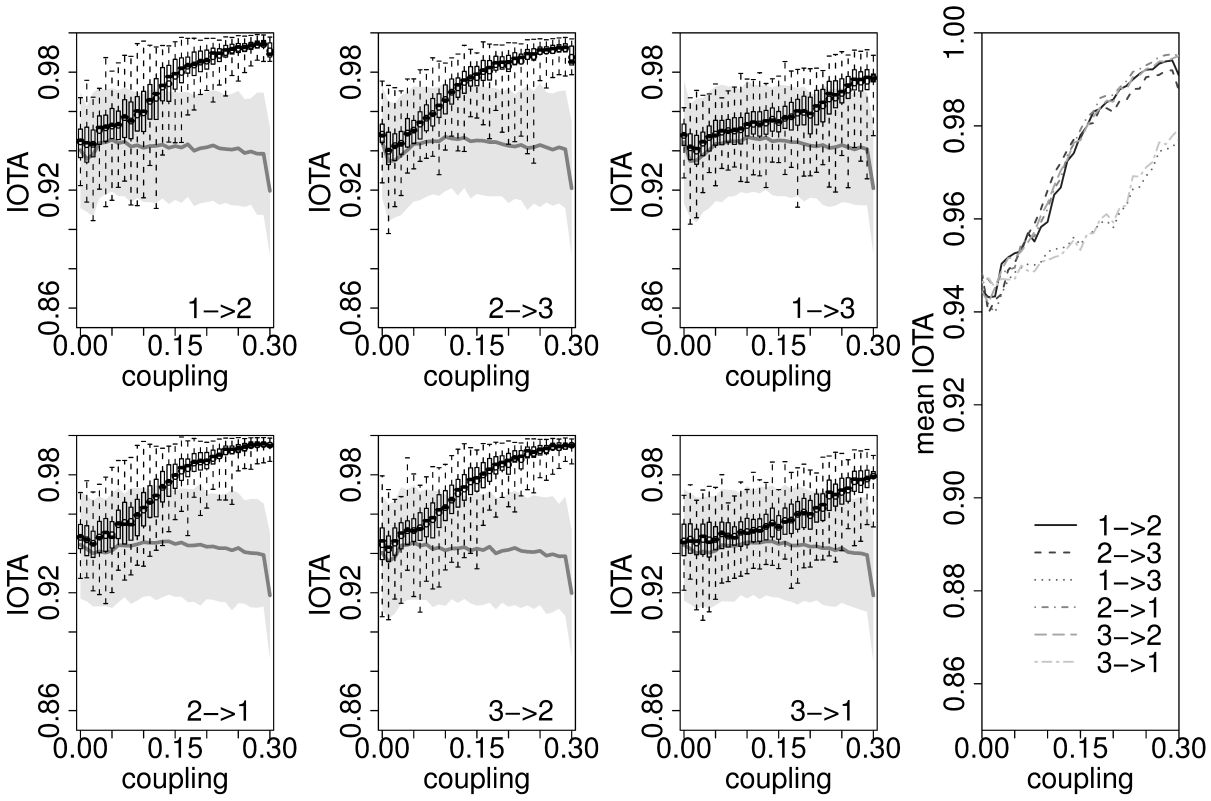

Figure 6.13: *IOTA* for Roessler oscillators (non-phase-coherent regime) coupled in a chain conformation with bidirectional coupling. The illustration is analogous to Fig. 6.10

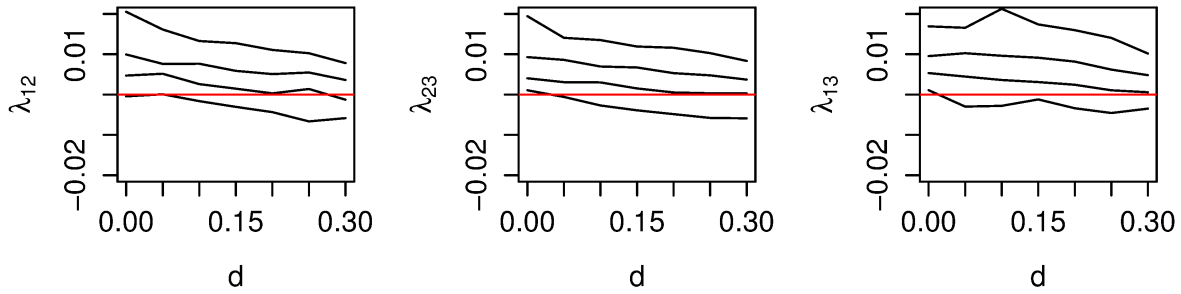

Figure 6.14: Lyapunov spectra for pairs of oscillators in the non-phase-coherent regime. Illustration as in Fig. 6.11

Comparison to available synchronization measures

Besides the standard association measures I discussed previously in this work (*e.g.*, correlation or mutual information), the couplings between chaotic oscillators are often inferred from more specific measures. For instance, in the phase-coherent regime these measures usually rely on the estimation of the phases $\Phi^{(k)}$ to characterize the degree of similarity.

There are few approaches that are designed to detect directional couplings between time series [RP01, OMWL08], however, most of the measures aim at quantifying the overall level of synchronization in a symmetric way. A common example is the phase synchronization index (mean phase coherence) [QQKKG02, SWD$^+$06, NRT$^+$10]

$$\mu_{PS}^{(k,l)} = \frac{1}{\tau}\sum_{t=1}^{\tau} e^{\Phi^{(k)}(t)-\Phi^{(l)}(t)} \tag{6.14}$$

those performance was studied in [SWD$^+$06] for a chain of coupled Roessler oscillators similar to the ones discussed above in Fig. 6.8 (a). The considered coupling scheme and coupling strengths are identical. Moreover, the dynamics of the coupled system is similar to the phase-coherent case considered above, where the parameters in [SWD$^+$06] correspond to $a = 0.15$, $b = 0.2$, $c = 10$, $\omega_1 = 0.99$, $\omega_2 = 1.03$, $\omega_3 = 1.01$ in Eq. (6.13) and additionally in [SWD$^+$06] Gaussian noise with standard deviation $\sigma = 1.5$ is added to the equations of the first component of each oscillator. Nonetheless, the phase synchronization index reaches values larger 0.5 (indicating a high probability of coupling) approximately for those coupling strengths where the values of $IOTA$ are also significant. However, the estimation of the phase is, in the general case, time-consuming and imprecise. Furthermore, there are several approaches to estimate the phase of a chaotic oscillator [PRK01] (*e.g.*, Poincare projection, displacement of the velocity, Hilbert- or Wavelettransform). However, for certain dynamical behavior of the system, these approaches may lead to very different results and, particularly when the chaotic oscillators are in the non-phase-coherent regime, the phase-based measures are not reliable. On the other hand, $IOTA$ reveals the links at similar coupling strengths with less computational effort, is applicable to rather short time series and works also in the non-phase-coherent regime.

Another approach to analyze the coupling between chaotic oscillators uses measures that are based on phase space reconstruction, *e.g.*, several (asymmetric) measures of nonlinear interdependence [QQKKG02], or recurrence based synchronization [NRT$^+$10]. Since these measures do not rely on the approximation of the phase they can be applied to study the non-phase-coherent regime. In [NRT$^+$10] the recurrence based synchronization

$$\mu_{RS}^{(k,l)} =< R_k(\epsilon,\tau), R_l(\epsilon,\tau) >, \tag{6.15}$$

the correlation of the probabilities of recurrence[4] R, was applied to investigate coupled Roessler oscillator in the non-phase-coherent regime which are arranged in a chain conformation identical to the scheme in Fig. 6.8 (a). The results obtained with the recurrence based synchronization in [NRT$^+$10] are directly comparable to those obtained with $IOTA$ here, since identical parameter

[4]The probabilities of recurrence can be estimated from the recurrence plot for threshold ϵ as the diagonal-wise calculated τ-recurrence rate

have been chosen for the oscillators in Eq. (6.13). However, the length of the investigated time series in [NRT$^+$10] was more than 15 times that of the length used to calculate $IOTA$. Nevertheless, $IOTA$ obtained significant values, at least with respect to the quartile, for very similar couplings strengths as the recurrence based synchronization measure, and for slightly stronger coupling the values of $IOTA$ are significant also with respect to its extreme values. The proper reconstruction of the phase space requires long time series and is strongly parameter dependent. $IOTA$, however, is independent on such estimations and thus excludes several error sources, requires less computational effort and works well also for rather short time series.

Influence of the available length of the data

In order to check the robustness of the obtained results, next, short time series of length 150 are considered and the influence of the number of available time points is exemplarily studied for the scenario of three coupled oscillators. First, the impact of the sampling is investigated. In that context, a coarse-grained version of the previously used trajectories is analyzed, where only every $10th$ point of the time series is regarded. This results in 100 time series consisted of 150 time points each (sampled at 2 Hz).

A reduction of the number of considered time points from 1500 to 150, while the same region of phase space is considered, does not much impair the inference of coupling in the phase-coherent regime, except for a slightly increased variance of $IOTA$ and a larger 2σ interval for the non significant values (Fig. 6.15, blue curves). In the non-phase-coherent regime the same two effects occur where particularly the broadening of the corridor is perceptibly, which requires larger coupling strength to obtain again significant values of $IOTA$ (Fig. 6.16, blue curves).

Next, the number of oscillations is reduced in addition. The system is again numerically integrated using the same Runge-Kutta scheme, but in this case time series of length 32000 sampled at 20 Hz are generated starting from random initial conditions. To get rid of transient behavior again the first 2000 time points are neglected. The rest of the trajectory is splitted into 100 pieces of length 300 (corresponding to approximately 2 oscillations). A coarse-grained version of the simulated trajectories is studied, where only every $2nd$ point is regarded which results in time series of length 150 sampled at 10 Hz.

In this case, a different region of the phase space is investigated and less of the dynamics of the system is represented by each of the short trajectories compared to the previous example. As a result, compared to the sampling at 2 Hz, larger coupling strengths are required in order to obtain significant values of $IOTA$ with a small variance (in particular in the non-phase-coherent regime shown in Fig. 6.16). However, the values of $IOTA$ for increased coupling strength behave in general in a very similar way as in the previous case. In particular the shape of the curves is unaffected (Fig. 6.15 and 6.16, black curves versus blue curves). On the other hand, the $IOTA$ values obtained for the randomized time series are lower if the sampling is 2 Hz instead of 10 Hz, *i.e*, the 2σ interval is shifted down rendering the values more significant. This is evident particularly in the non-phase-coherent regime.

In summary this means that the boundary value for which $IOTA$ is first regarded as significant mainly depends on the number of time points per oscillation, while the width of the 2σ interval and the variance of the $IOTA$ values are basically determined by the length of the time series.

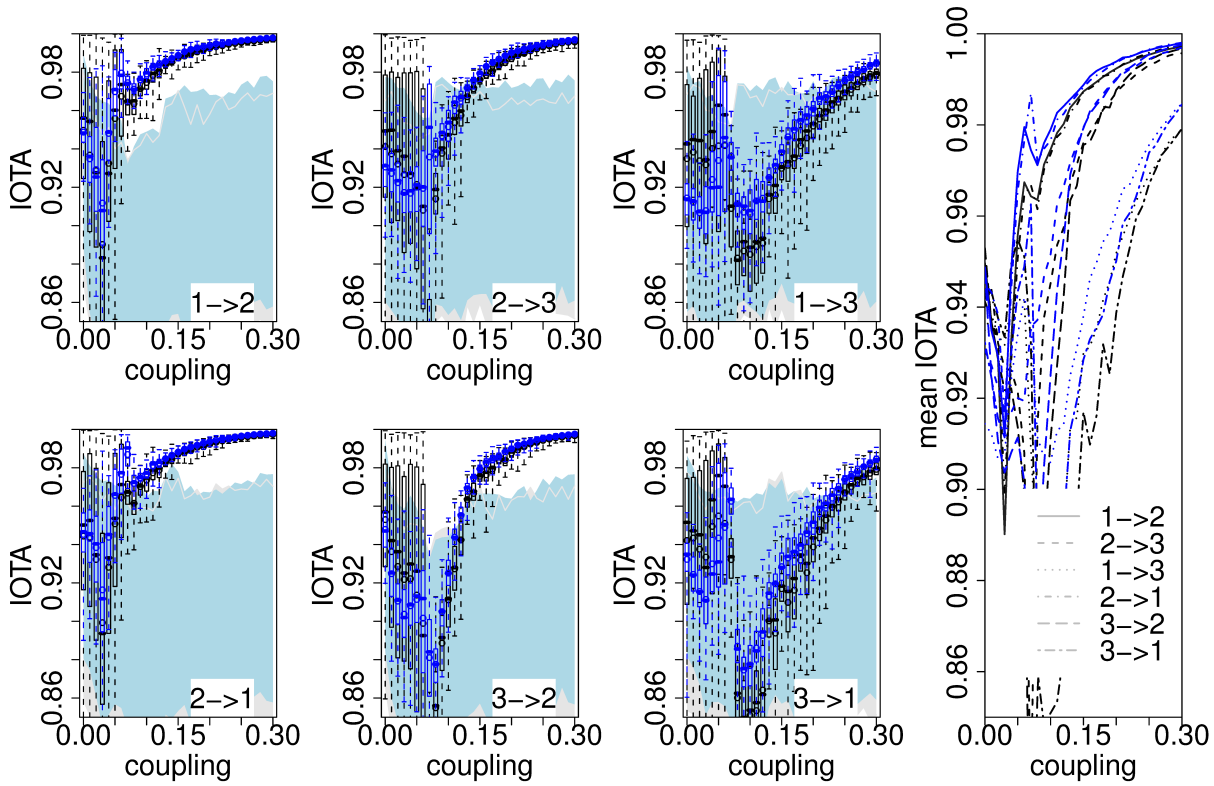

Figure 6.15: *IOTA* for short trajectories of 3 phase-coherent, coupled Roessler oscillators (150 time points). The time series capture a different number of oscillations; blue: 11 , black: 2. The illustration is the same as in Fig. 6.10.

Unidirectional and mixed coupling

In what follows, the chain of 3 coupled oscillators, with time series of length 1500 are revisited. This time the bidirectional links are (partially) replaced with unidirectional ones, thus, representing different typical network motifs (as shown in Fig. 6.9 (a)–(c)).

First, the phase-coherent regime is studied for the different coupling situations, where the coupling d in coupling matrix D is varied from 0.0 to 0.3 in steps of 0.01 for the simulation.

In case (a) oscillator 2 is a common driver of 1 and 3. Fig. 6.17 illustrates that the coupling between oscillator 1 and 2 becomes significant approximately for $d = 0.24$, and the one between 2 and 3 around $d = 0.07$; for the indirect coupling between oscillator 1 and 3 $IOTA$ is significant only for $d > 0.27$. Same as in the bidirectional scenario before, $IOTA$ obtains significant values for the direct links starting at coupling strengths around the estimated onset of synchronization. The directionality of the links is not inferable by means of absolute values of $IOTA$ (since there is no delay between the drive and the response system). However, the randomization of the time series results in larger variations of $IOTA$ for the correct coupling direction compared to the false one. This indicates the links $2 \rightarrow 1$ and $2 \rightarrow 3$ to be unidirectional, while $1 - 3$ is

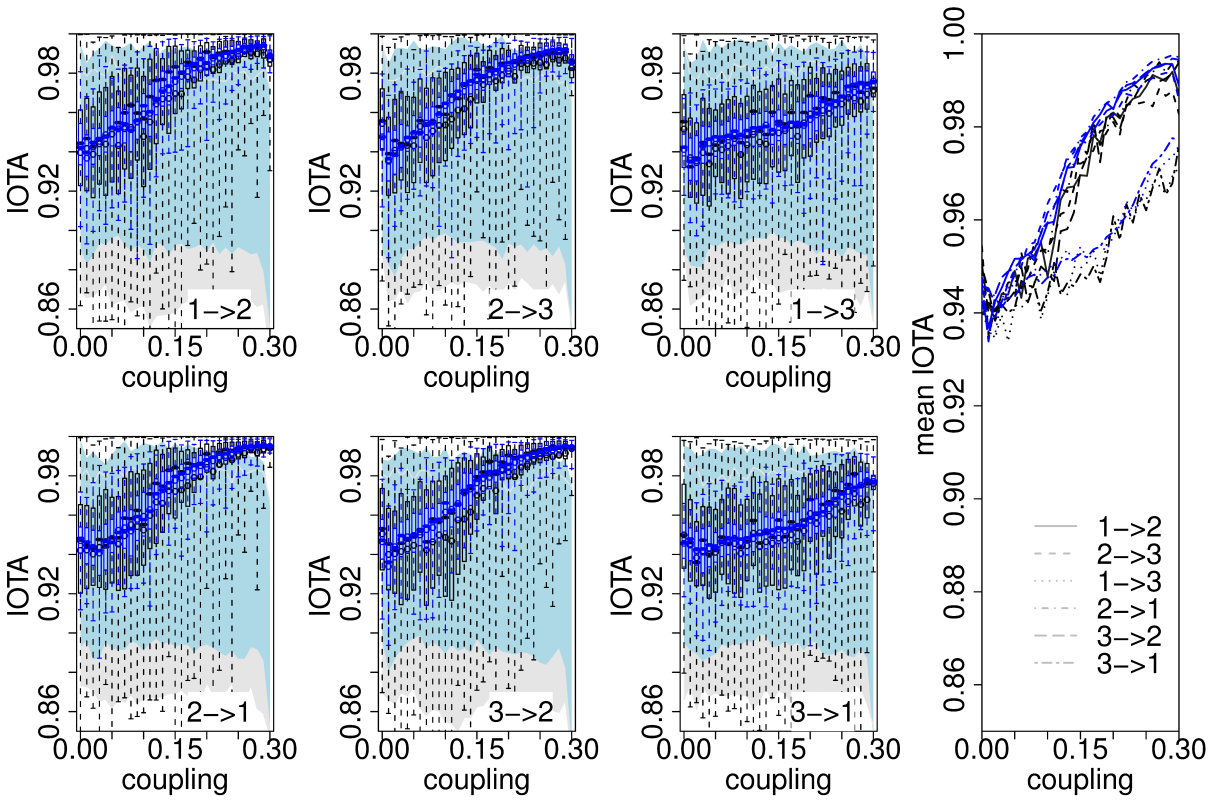

Figure 6.16: $IOTA$ for short trajectories of 3 non-phase-coherent, coupled Roessler oscillators (150 time points). The time series capture a different number of oscillations; blue: 11 , black: 2. The illustration is the same as in Fig. 6.10.

more likely to be bidirectional. The partial measure indicates the direction $2 \rightarrow 3$ to be the most likely one, while for the other links a distinction between direct and indirect links is barely possible (Fig. 6.18 (a)).

In case (b) a cascade is realized where oscillator 1 is driving 2 which on the other hand drives 3. The related values of $IOTA$ are shown in Fig. 6.19. The direct links obtain significant values of the measure ($d > 0.08$ for 1 and 2, $d > 0.16$ for 2 and 3) around the coupling strength that is related to the onset of synchronization, and the indirect links become significant only at larger coupling strengths ($d > 0.24$ for 1 and 3). Moreover, the values of $IOTA$ are much lower for the indirect links between oscillator 1 and 3 than for the other possible links. While the directionality of the coupling is not inferable from the absolute values of $IOTA$, the variations of $IOTA$ for the randomized time series indicate $1 \rightarrow 2$ correctely to be a unidirectional link. Additionally, the pairwise measure suggests the coupling between 2 and 3 as bidirectional link and the one between 1 and 3 as unidirectional link directed from 1 to 3. The partial measure (Fig. 6.18 (b)), on the other hand, indicates that not only the link $1 \rightarrow 3$ must be preferred over $3 \rightarrow 1$, but also $2 \rightarrow 3$ over $3 \rightarrow 2$.

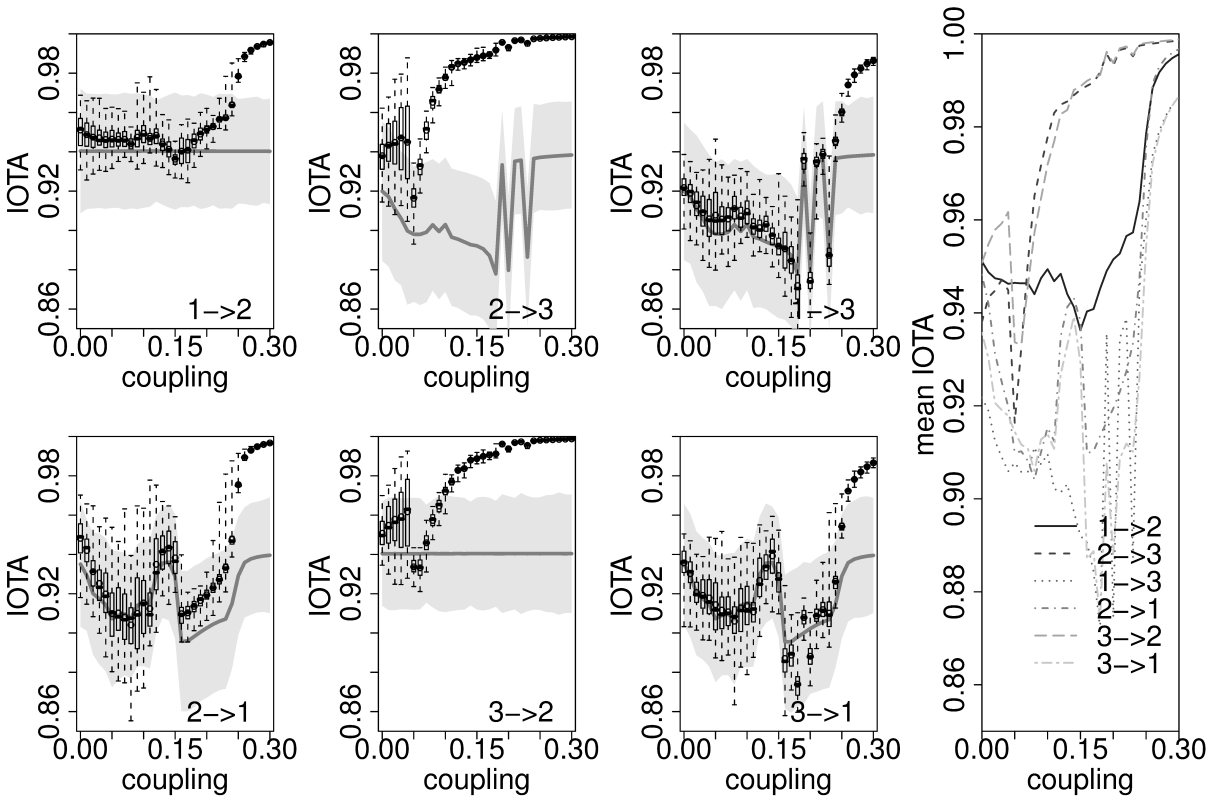

Figure 6.17: *IOTA* for Roessler oscillators (phase-coherent regime) coupled in a chain conformation with unidirectional (common driver) coupling. Illustration analogous to Fig. 6.10.

Finally, in case (c) a mixture of uni- and bidirectional links is considered. From Fig. 6.20 it becomes apparent that the inference of this coupling situation is more problematic than in the previous cases. While the bidirectional link between oscillator 2 and 3 is indicated correctly even for very low coupling strength (d=0.04), on the other hand, the correct unidirectional link $2 \rightarrow 1$ is barely detectable. The coupling between oscillator 1 and 2 becomes significant for coupling strengths of approximately $d = 0.25$, however, the indirect link between 1 and 3 only little later for coupling strengths of approximately $d = 0.29$. Additionally, these 4 links obtain similar values of $IOTA$.

Furthermore, while the directionality is again not inferable from the absolute values of $IOTA$, nevertheless, the variations of $IOTA$ for the randomization of the time series suggest to prefer the link $3 \rightarrow 2$ over $2 \rightarrow 3$, $2 \rightarrow 1$ over $1 \rightarrow 2$, and $3 \rightarrow 1$ over $1 \rightarrow 3$. It is not obvious from the pairwise measure which of these links are actually unidirectional ones. The partial measure, on the other hand, indicates correctly the bidirectional link between 2 and 3, while the other links (at least for coupling strengths $d > 0.25$) are more likely to be unidirectional. However, a distinction between direct and indirect links is barely possible with the partial measure, although

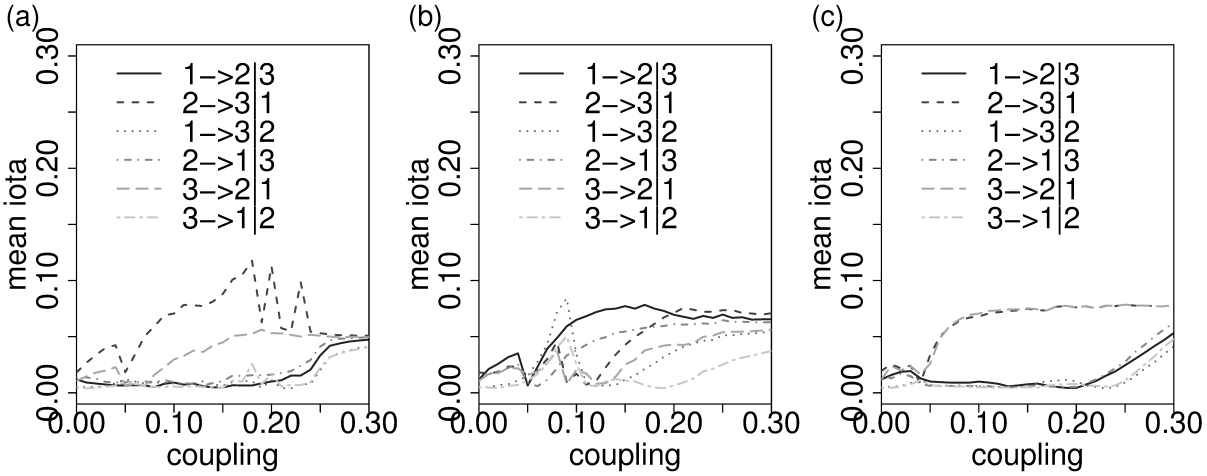

Figure 6.18: Mean of partial $IOTA$ for 3 phase-coherent Roessler oscillators in a chain assembly with partially unidirectional coupling: (a) common driver, (b) cascade, (c) mixture

for larger coupling strengths (where the pairwise measure is significant for all links) the three largest values of the partial $IOTA$ are indeed obtained for the three direct links (Fig. 6.18 (c)).

If the coupled oscillators are not phase-coherent, but in the Funnel regime for none of the unidirectional links (Fig. 6.21–6.23) significant values of $IOTA$ are obtained within the considered range of coupling strengths, however, the bidirectional link in case (c), Fig. 6.23, shows significant values of the measure for $d > 0.15$.

As a result, the direct and indirect unidirectional links are not distinguishable here. Nevertheless, some trends show up for the unidirectional links in the different network motifs. In Fig. 6.21 the resulting $IOTA$ values for case (a) are shown. $IOTA$ tends to obtain higher values for increased coupling strengths, and at $d \approx 0.25$ for the directly coupled oscillator pairs the values, with respect to the quartile, become significantly different from those obtained for the randomized time series. However, the variance of $IOTA$ is still large and, thus, the unidirectional links together with the correct directionality are not inferable (both directions will be kept). Additionally, the values of $IOTA$ for the coupling between oscillator 1 and 3 are slightly lower than for the other links suggesting that these links between 1 and 3 might be indirect.

Almost the same situation occurs in Fig. 6.22 for case (b), however, the values of $IOTA$ for the indirect link are lower than in the previous case. This is most likely the result of an implicit delay in the regulation from 1 to 3, since the driving is mediated by oscillator 2. Thus, the distinction between direct and indirect links is feasible in this case.

Eventually, the partial measure does not provide additional information on the coupling situation for the unidirectionally coupled oscillators in the non-phase-coherent regime (Fig. 6.24) within the considered range of coupling strengths.

Furthermore, it shows up that the significance of the results of the coupling analysis as well as the variations of the $IOTA$ values strongly depend on the particular number of oscillations captured by the time series, where more oscillations result in less variation and allow for a better

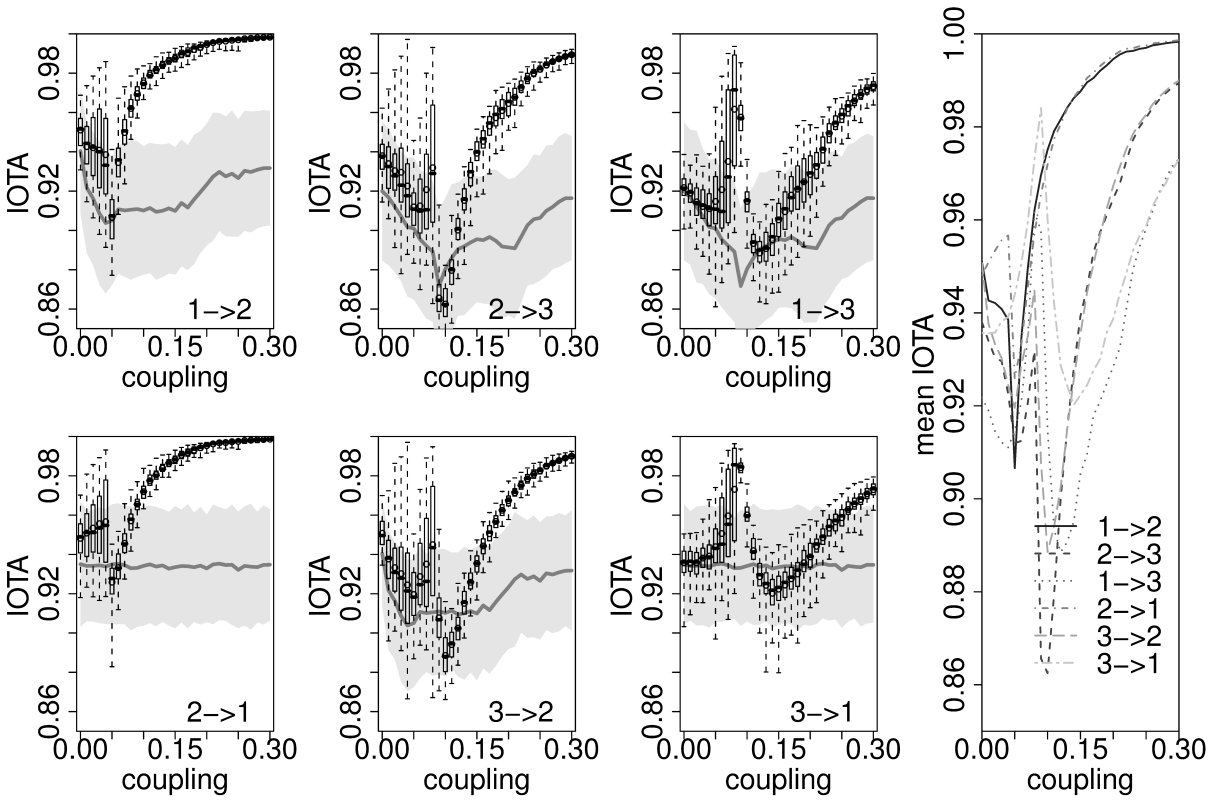

Figure 6.19: *IOTA* for Roessler oscillators (phase-coherent regime) coupled in a chain conformation with unidirectional (cascade driver) coupling. Illustration analogous to Fig. 6.10.

distinction between bi- and unidirectional links, but also require larger values of *IOTA* obtain significant results. In particular, this behavior is analyzed for the 3 partially unidirectional coupled Roessler oscillators (Fig. 6.9 (c)) in the phase-coherent regime using short time series with 150 time points, which include 2 (sampling at 10 Hz) and 25 (sampling at 1 Hz) oscillations, respectively.

In Fig. 6.25 the interval which indicates not significant values of *IOTA* ends at higher values of *IOTA* in the case of 25 oscillations than in the case of 2 oscillations, which is in accordance with the observation made for the bidirectional coupling. Thus, in the latter case, even for low coupling or indirect links the measure can obtain significant values. However, the variance of the obtained *IOTA* values is also much higher in that case. Additionally, in none of both cases the directionality can be inferred from the absolute values of *IOTA*. Nevertheless, the distinction between uni- and bidirectional links (in the phase-coherent case) is possible from the variation of the results for the randomized time series and the coupling strength for which significant values are observed for the first time. In particular, the difference in that coupling strength is increased if the time series capture more oscillations.

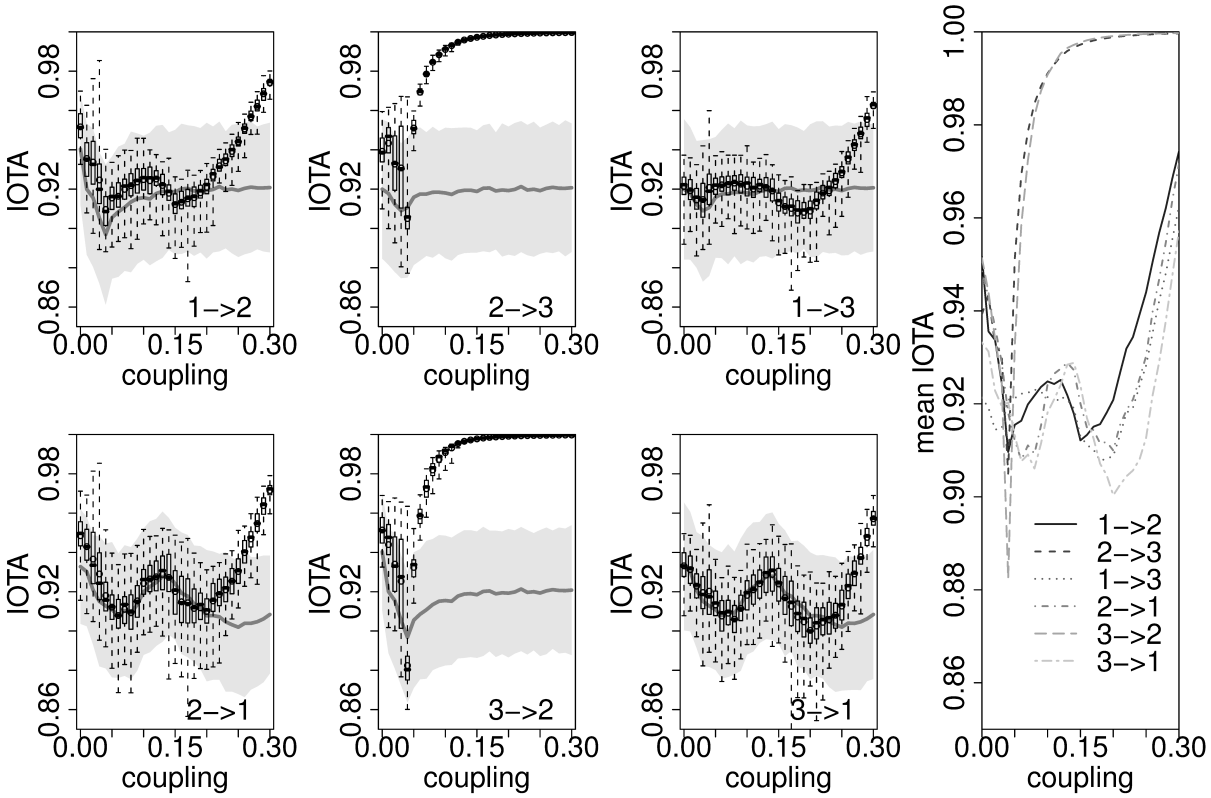

Figure 6.20: *IOTA* values for Roessler oscillators (phase-coherent regime) coupled in a chain conformation with unidirectional coupling from 2 to 1 and bidirectional one between 2 and 3. Illustration analogous to Fig. 6.10.

Influence of the network size

Usually real systems consist of more than 3 subsystems which may complicate the inference of coupling, since $IOTA$ (same as other measures) treats only pairs and triplets of subsystems. Hence, I expand the initial analysis on bidirectionally coupled Roessler oscillators to a scenario with 4 oscillators arranged as a star. In this scenario subsystem 3 is bidirectionally coupled to each of the other subsystems (Fig. 6.8 (b)). Again both dynamical regimes, the spiral and the Funnel one, are investigated where the coupling strength d in coupling matrix D (Fig. 6.8 (b)) is varied from 0.0 to 0.3 in the spiral and up to 0.4 in Funnel regime (in steps of 0.01).

For spiral chaos the boxplots in Fig. 6.26 and 6.27 illustrate that the oscillator pair $1-3$ obtains high values of $IOTA$ even for very low couplings ($d > 0.02$) where for $d > 0.14$ additionally the variance is low. For the latter coupling value the pairs $2-3$, $3-4$ and $2-4$ also obtain large values of $IOTA$ with low variance. Furthermore, the measure's values for the pairs $1-2$ and $1-4$ become significant only for slightly larger coupling strengths. However, the values for these pairs are much lower and the variance is larger, which suggests that $1-2$ and $1-4$ are indirect links.

On the other hand, it is not obvious from the pairwise measure which links from the triplet

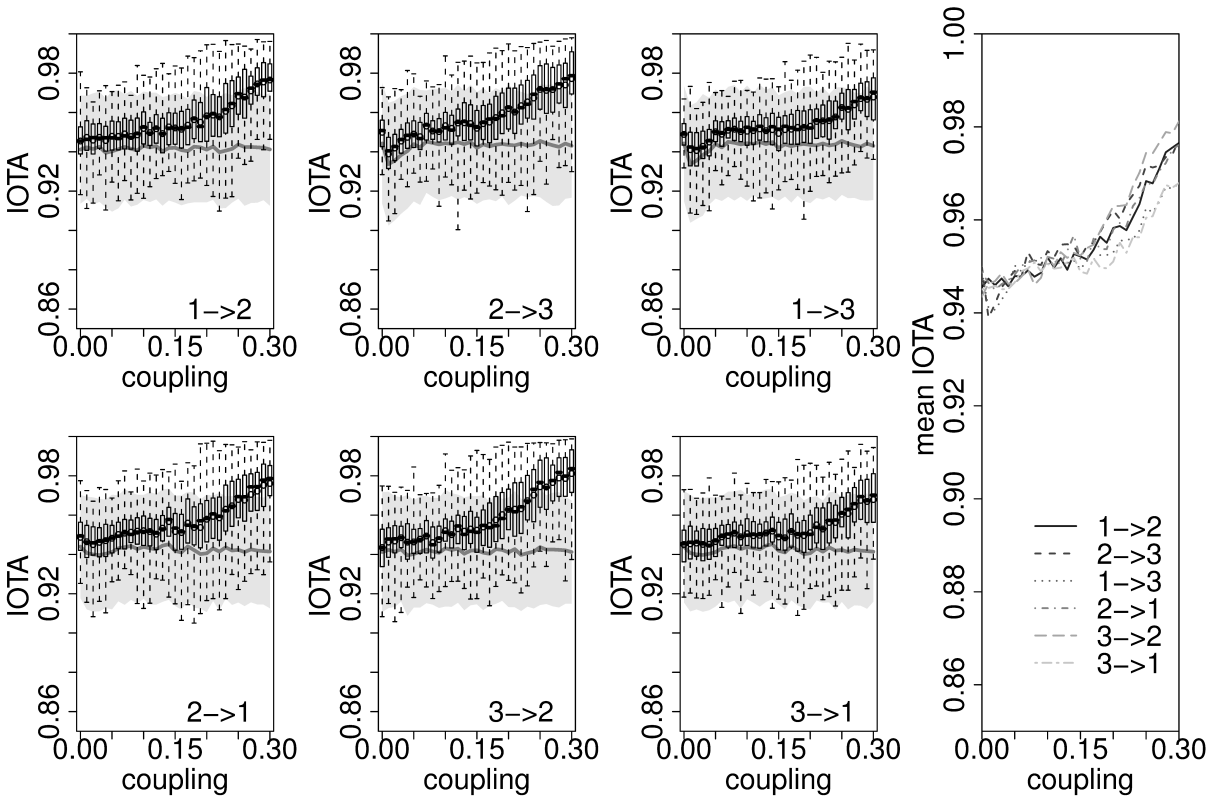

Figure 6.21: $IOTA$ for 3 coupled non-phase-coherent oscillators (common driver). Illustration analog to Fig. 6.17, but for the non-phase-coherent regime.

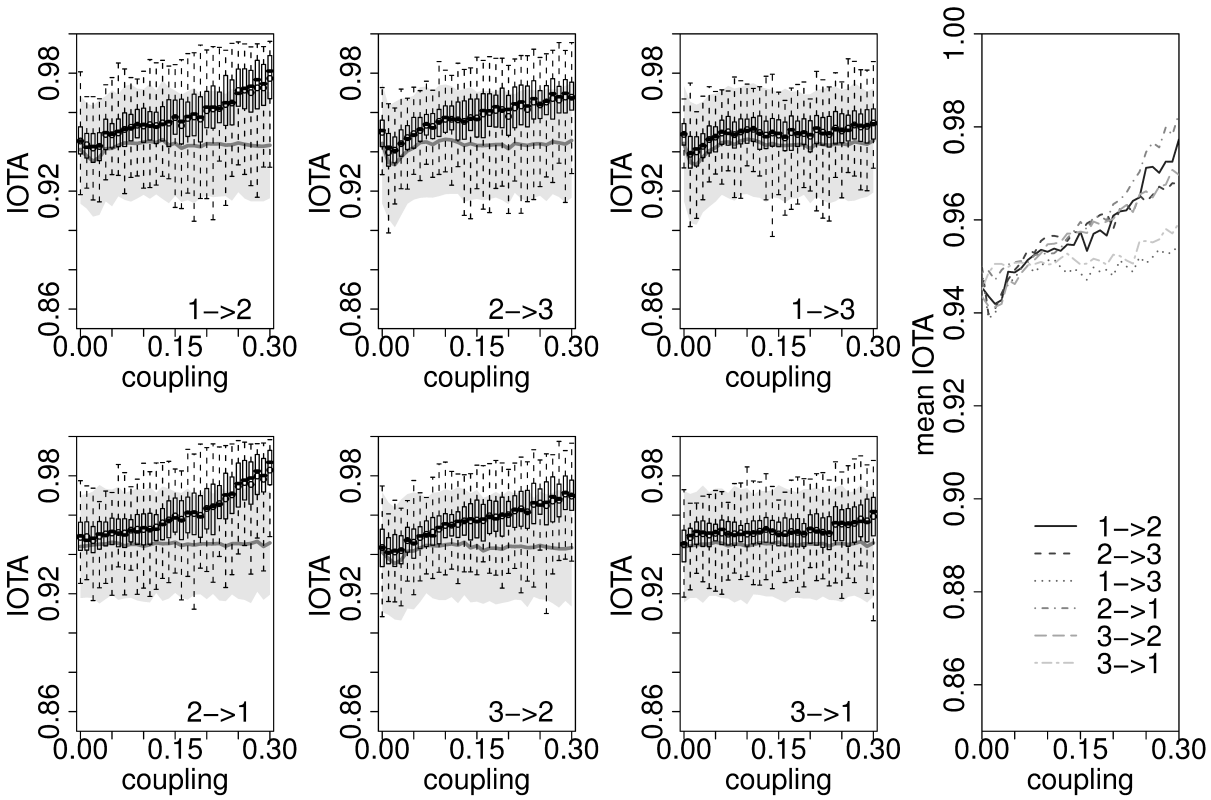

Figure 6.22: $IOTA$ for 3 coupled non-phase-coherent oscillators (cascade driver). Illustration analog to Fig. 6.19, but for the non-phase-coherent regime.

$2-3$, $3-4$ and $2-4$ are superfluous. In particular, the indirect link $2-4$ cannot be excluded here, which is, however, again in accordance with the estimated pairwise Lyapunov spectra (Fig. 6.28). The onset of phase synchronization between oscillator 2 and 4 occurs at very low coupling strength leading to a high similarity of the two time series.

All oscillator pairs obtain very similar values of $IOTA$ for both directions. Additionally, $IOTA's$ variance for the randomized series shows no tendency for unidirectional links. Thus, it is most likely that the inferred coupling scheme includes only bidirectional links.

Furthermore, even with the partial measure, not all the indirect links are directly inferred. For intermediate coupling strengths the partial measure (Fig. 6.29) indicates correctly that $1-2$ and $1-4$ are indirect. In contrast, for the triplet $2-3$, $3-4$ and $2-4$ none of the pairs obtains significantly lower values than the other ones. Hence, it is difficult to judge which are the superfluous links in that case. However, for several coupling strengths the partial measure suggests a unidirectional coupling from $3 \to 2$, $2 \to 4$ and $4 \to 3$, where the values of the partial $IOTA$ are in general closer together for $2 \to 4$ and $4 \to 2$ than in the case of $3 \to 2$ and $2 \to 3$ or $3 \to 4$ and $4 \to 3$. Hence, choosing the link $2-4$ to be the superfluous one is most likely, since it meets best the observation that all links should be bidirectional (which is suggested by

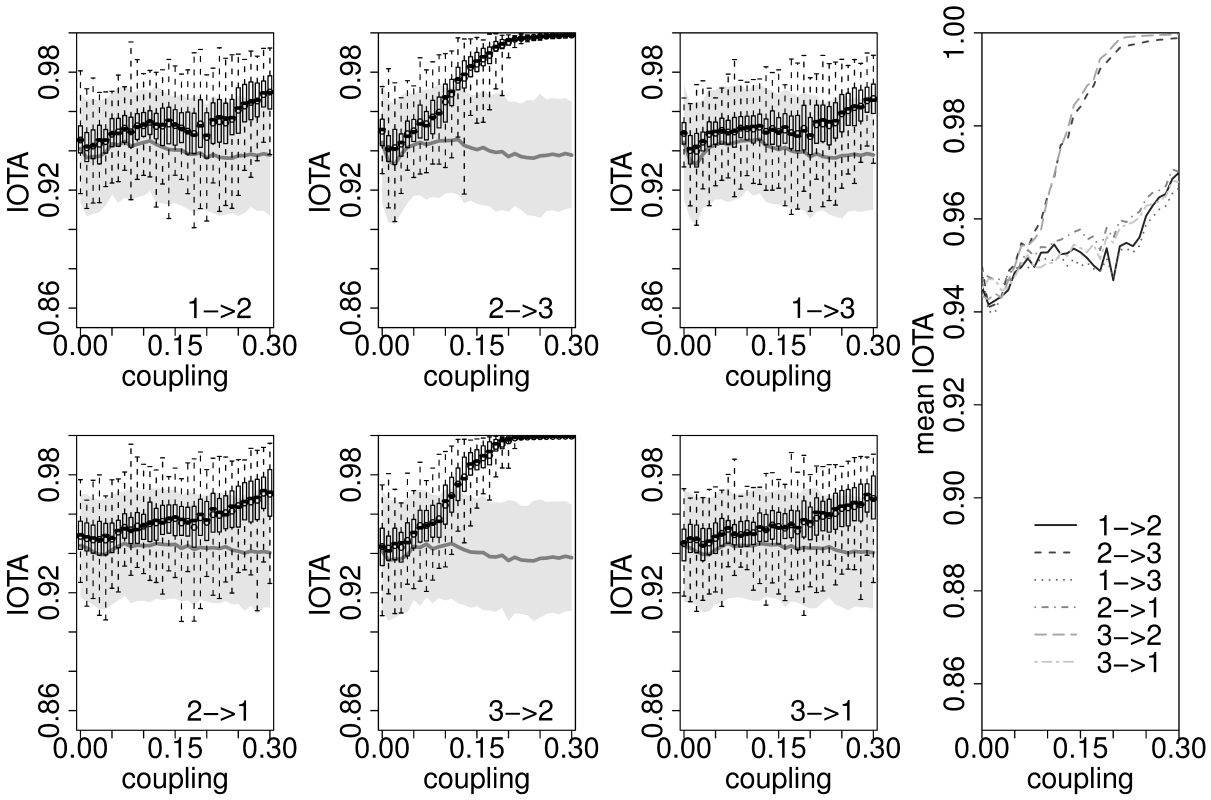

Figure 6.23: *IOTA* for 3 coupled non-phase-coherent oscillators (mixture). Illustration analog to Fig. 6.20, but for the non-phase-coherent regime.

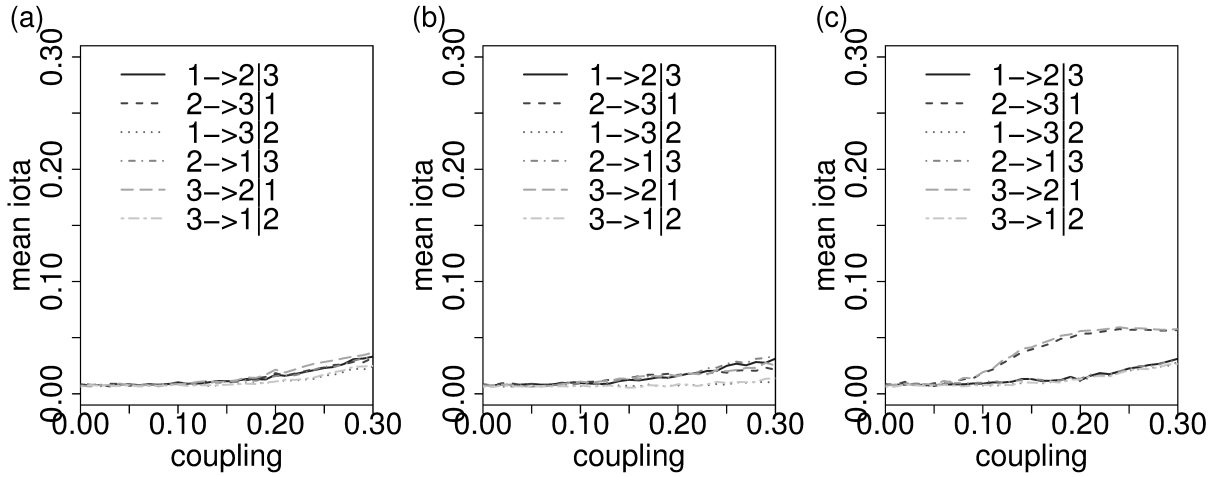

Figure 6.24: Partial *IOTA* for 3 unidirectionally coupled non-phase-coherent oscillators. Illustration analog to Fig. 6.18, but for the non-phase-coherent regime. (a) common driver, (b) cascade, (c) mixture

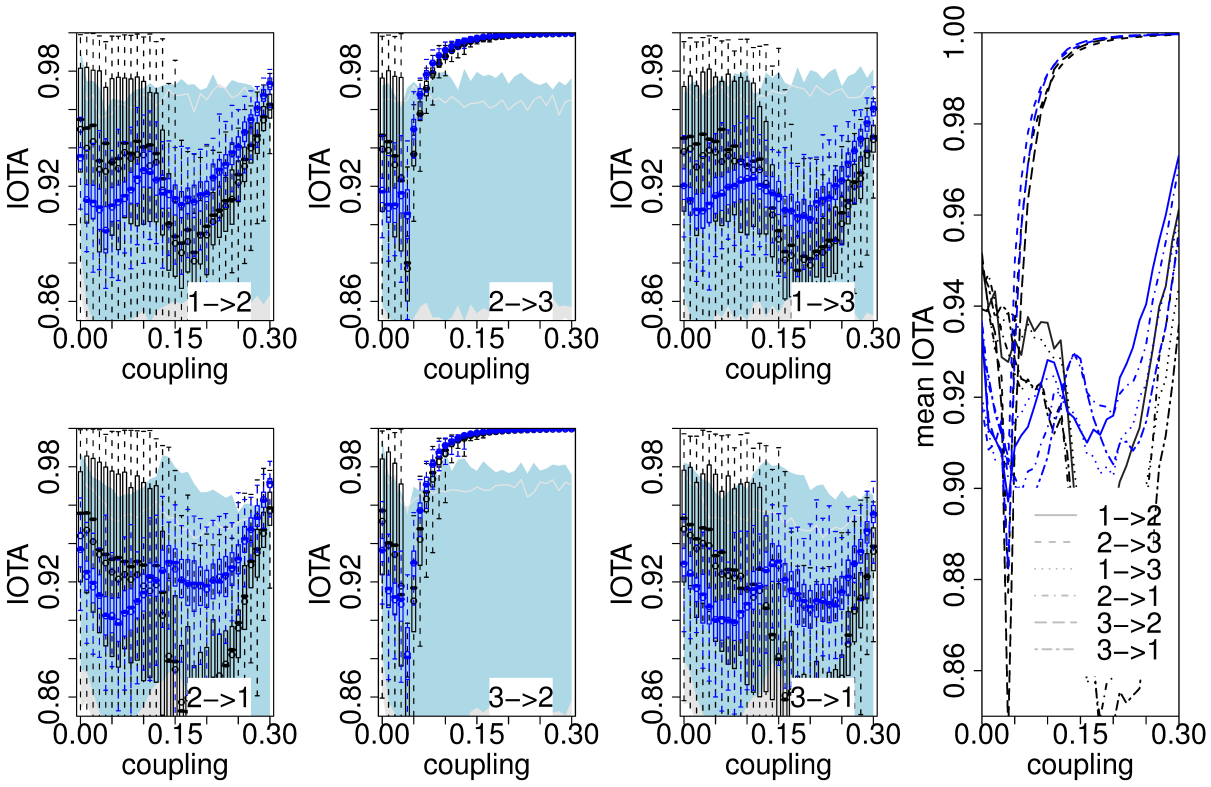

Figure 6.25: *IOTA* for Roessler oscillators (phase-coherent regime) coupled in a chain conformation with uni- and bidirectional coupling. Estimation from short time series (150 time points) which capture 2 (black curves) or 25 oscillations (blue curves). The illustrations is the same as in Fig. 6.20.

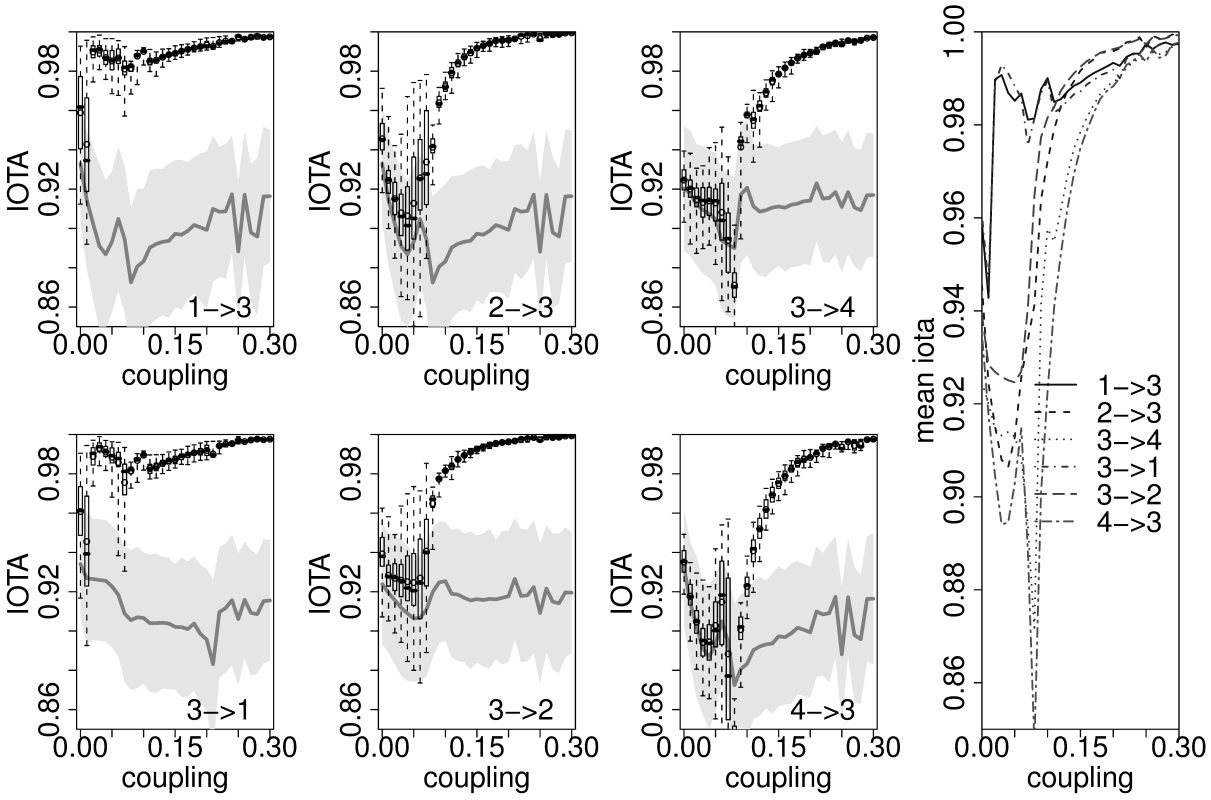

Figure 6.26: *IOTA* for Roessler oscillators (phase-coherent regime) coupled in a star conformation with bidirectional coupling. The illustration is analogous to Fig. 6.10

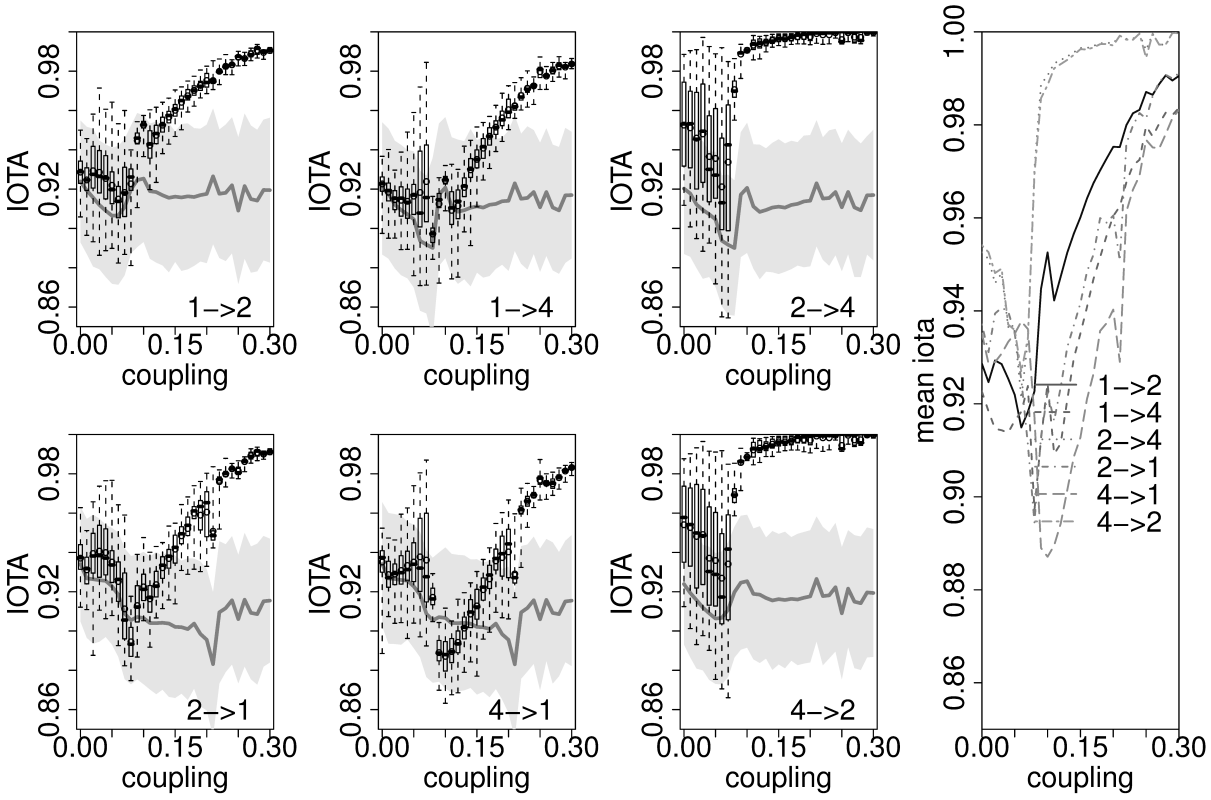

Figure 6.27: *IOTA* for Roessler oscillators (phase-coherent regime) coupled in a star conformation with bidirectional coupling. Illustration as in Fig. 6.10.

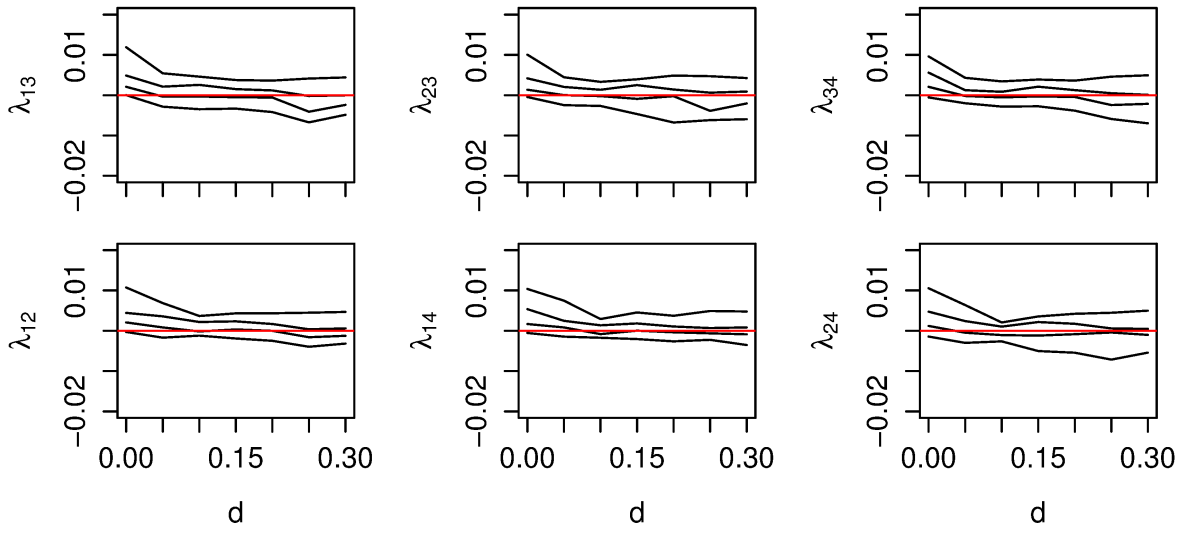

Figure 6.28: Lyapunov spectra for pairs of oscillators in the phase-coherent regime. Illustration analogous to Fig. 6.11, but for a system of 4 Roessler oscillators coupled in a star assembly.

the pairwise measure).

In contrast to the previous results, in the case of Funnel chaos the coupling situation is more obvious for the scenario of four coupled oscillators (Fig. 6.30 and 6.31): For coupling strengths larger $d = 0.3$ significant values of $IOTA$ are obtained for all direct links, while for the indirect links the variance of the measure is higher and the values are barely significant. Furthermore, from the mean values of $IOTA$ the direct and indirect links are well distinguishable over a wide range of coupling strengths.

The Lyapunov spectra (Fig. 6.32) indicate that all subsystems become synchronized at a coupling strength around $d = 0.3$ in the considered coupling scenario and dynamical regime. This further strengthens the observation that the coupling of chaotic oscillators can be inferred with $IOTA$ from the time series of one observable if the oscillators are close to synchronization.

The partial $IOTA$ measure (Fig. 6.33) confirms the coupling structure which is inferred from the pairwise measure.

However, for coupling strengths larger $d = 0.4$ the synchronization leads to almost identical time series for all oscillators, thus, the direct and indirect links are barely distinguishable anymore from the pairwise and the partial measure.

General findings

In summary, $IOTA$ has been proven a valuable tool for inferring the coupling between chaotic subsystems if only short time-resolved measurements are available. I showed that in general the measure is capable to infer couplings of chaotic Roessler oscillators in different dynamical

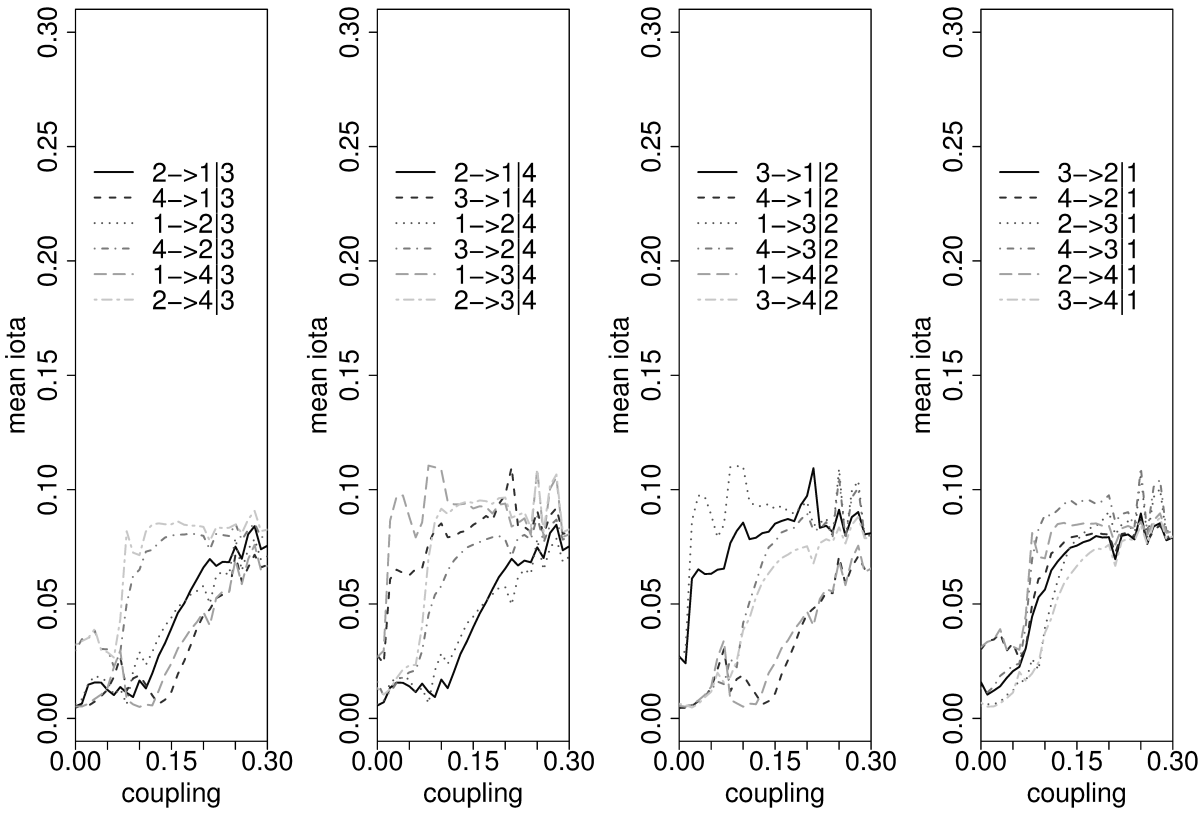

Figure 6.29: Mean of partial $IOTA$ over 100 trajectories for a star of 4 bidirectionally coupled Roessler oscillators in phase-coherent regime.

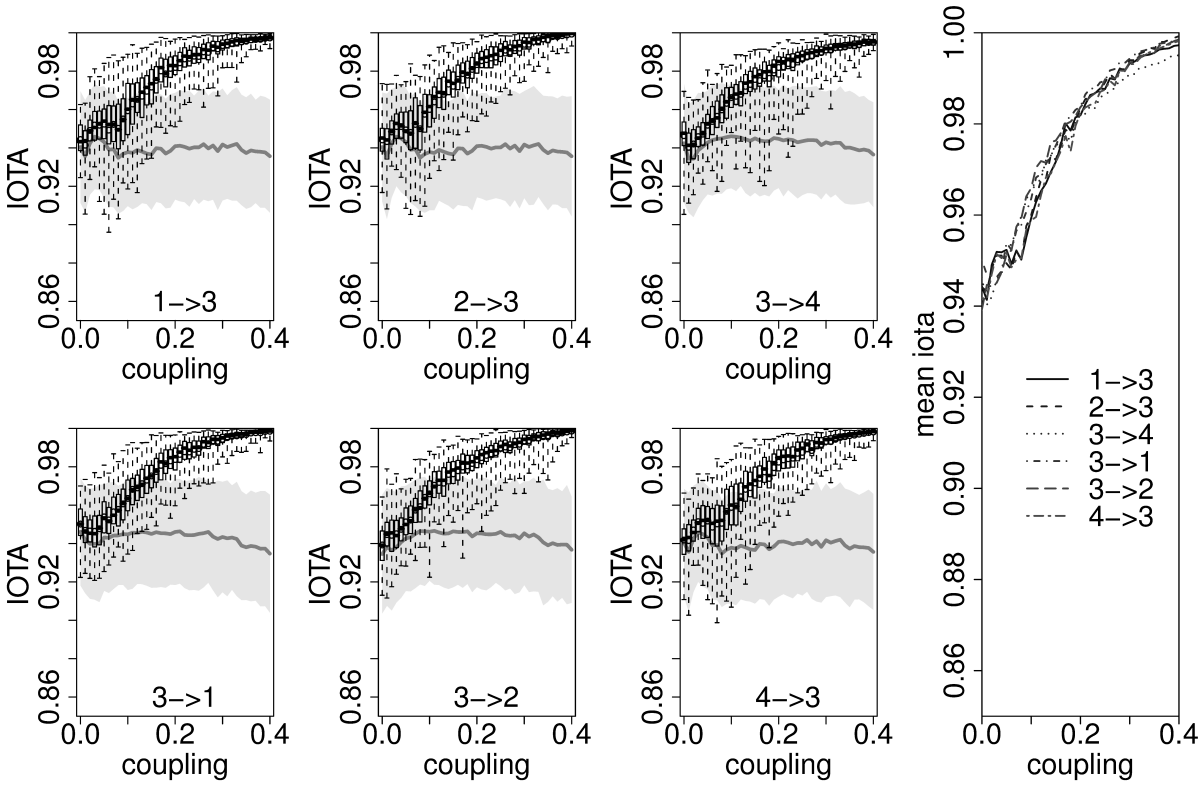

Figure 6.30: *IOTA* for different pairs of 4 bidirectionally coupled non-phase-coherent oscillators. Same illustration as Fig. 6.26, but for the non-phase-coherent regime.

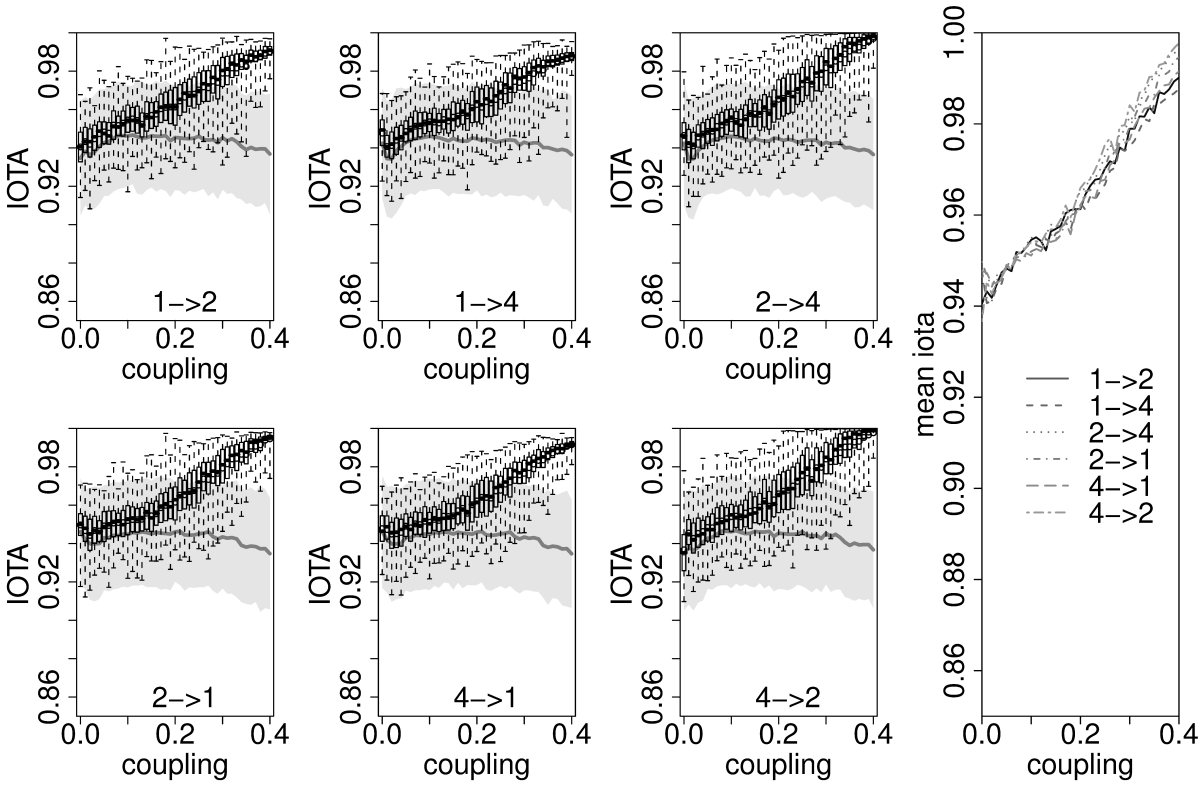

Figure 6.31: *IOTA* for 4 bidirectionally coupled non-phase-coherent oscillators. Illustration analogous to Fig. 6.30, but for different pairs.

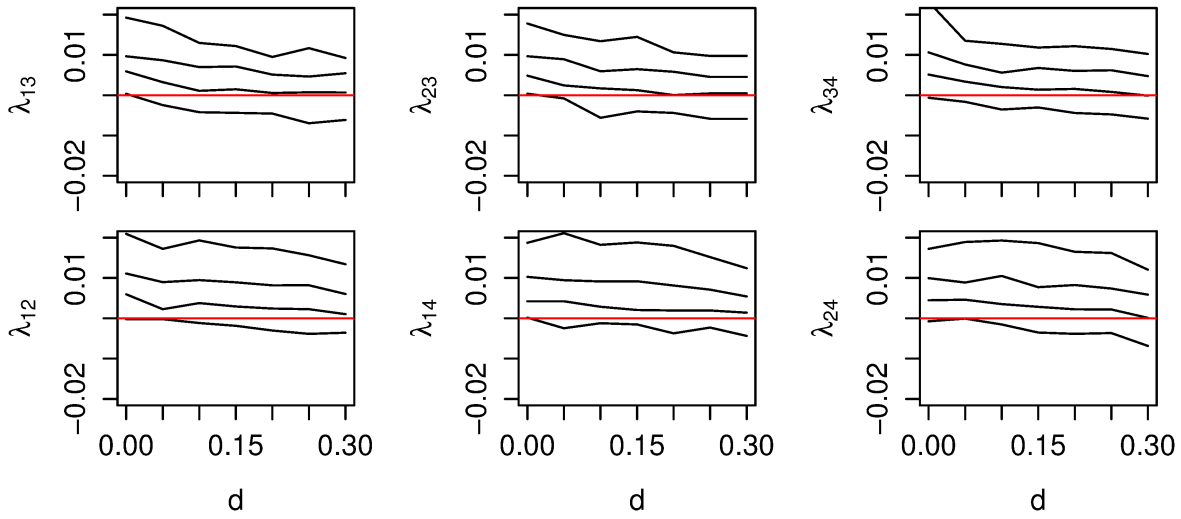

Figure 6.32: Lyapunov spectra for pairs of oscillators in the non-phase-coherent regime. Illustration as in Fig. 6.28.

regimes, where the non-phase-coherent regime necessitates stronger coupling than the phase-coherent one. If the coupling strength is to low the decision whether there is coupling or not will strongly depend on the trajectory's position in the phase space.

In both dynamical regimes, the necessary coupling strengths are in accordance with results obtained from other available measures applied to longer time series in previous studies. However, *IOTA* usually requires less time points and disclaims the approximation of the phase space.

Moreover, I observed that stronger coupling strengths are needed to obtain significant results for unidirectional coupling compared to bidirectional one. Hence, for coupling schemes including both types of coupling situations the unidirectional links might be hidden.

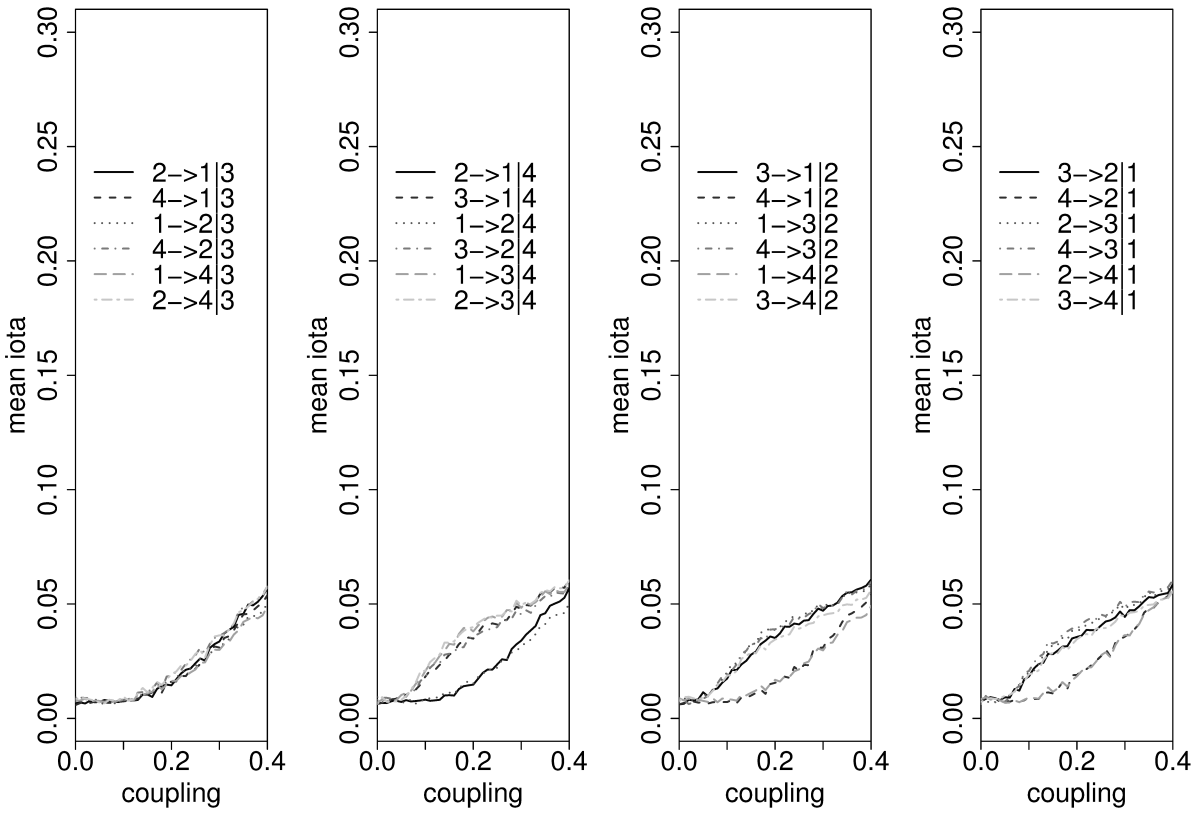

Figure 6.33: Partial $IOTA$ for 4 bidirectionally coupled non-phase-coherent oscillators. Illustration analog to Fig. 6.29, but for non-phase-coherent regime.

7 IOTA for reconstructing gene regulatory networks (GRN's)

Next, the reconstruction efficiency of the basic relevance network algorithm using $IOTA$ as association measure is investigated and compared to that of Kendall's τ. For this purpose, GRN's of different sizes and slightly different topology (*i.e.*, different dynamics) and the corresponding time series of gene expression are generated and analyzed. Furthermore, $IOTA$ is applied on experimentally obtained gene expression data and the reconstructed regulatory network is discussed.

7.1 Reconstructing GRN's from synthetic time-resolved data

While the previous numerical study was limited to investigate the properties of $IOTA$ for small network modules with 3 to 7 nodes, in real world applications the reconstruction of such small systems from short time series is unusual. In most of the cases it is necessary to determine the interrelations between nodes in large-scale networks of hundreds and thousands of subsystems. However, as the number of interacting subsystems increases, the collection and assessment of data is impaired. Thus, less data is available for the process of network reconstruction. Microarrays, as a prominent example, measure the expression of hundreds of genes (*i.e.*, the concentration of gene products such as mRNA) at few selected time points. Hence next, the capabilities of $IOTA$ will be further investigated for the reconstruction of GRN's, using synthetic data sets as a starting point. For this purpose the two, well-defined regulatory networks of the bacterium *E. coli* and the baking yeast *S. cerevisiae* are revisited to evaluate the ability of $IOTA$ to solve the network inference problem. Subnetworks and time series are generated using $SynTRen$ as described previously in this work (Section 1.3).

7.1.1 Dependence on the length of the time series

First, the (pairwise) measure is applied to investigate the dependence of the reconstruction on the length of the time series (within a realistic range) for a rather small subnetwork of *E. coli*'s GRN. More precisely, I analyze a network of 60 genes (representing the nodes) with 62 unidirectional links, 3 of which are autoregulatory ones. The dynamics of each node (gene) is governed by Michaelis-Menten and Hill kinetics. Both, deterministic and stochastic (noise level 0.01) time series, are investigated. They consist of 10, 30, 50, and 70 time points each, corresponding to measurement data. Figure 7.1 shows this network of 60 genes of *E. coli* and the corresponding simulated time series of 70 time points obtained from $SynTReN$.

Similarly to Section 6.1.1, the values obtained with the different measures are stored in a matrix $I = \mu(y^{(k)}, y^{(l)})$, where μ is either μ_ι, $\mu_{\tilde{\iota}}$, μ_{b_ι} or $|\mu_K|$ (in case of Kendall's τ). Next, in

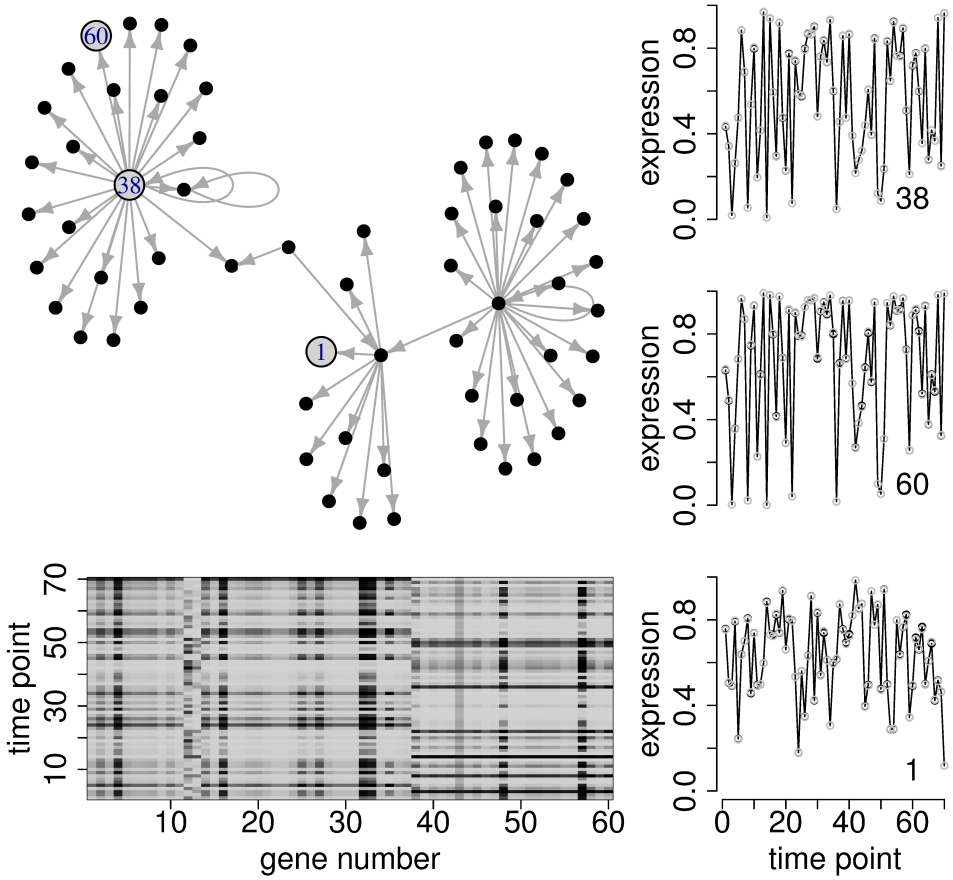

Figure 7.1: Illustration of a network obtained as a subnetwork of the GRN of *E. coli*. The time course of three genes (marked in the network) is shown on the right. In the lower left panel the time series of all genes in the network are visualized. Light indicates high concentration of gene products, while dark indicates absence of the gene product.

order to determine how well the measures distinguish actual coupling from random similarities, for each length of the time series the fraction of true links associated with the largest values of the corresponding measure is calculated

$$f_c = \frac{I_s^{link}}{I_s}. \tag{7.1}$$

Here, I_s is the number of entries of I larger than a threshold value s, whereas I_s^{link} corresponds to the part of I_s which coincides with actual coupling. To define the threshold value, first the amount c of true links in the network is determined. Next, all entries of the matrix I are ranked in decreasing order. Finally, the entry with rank c defines the threshold s.

Furthermore, the fraction of links where the true coupling direction could be inferred, f_d, is

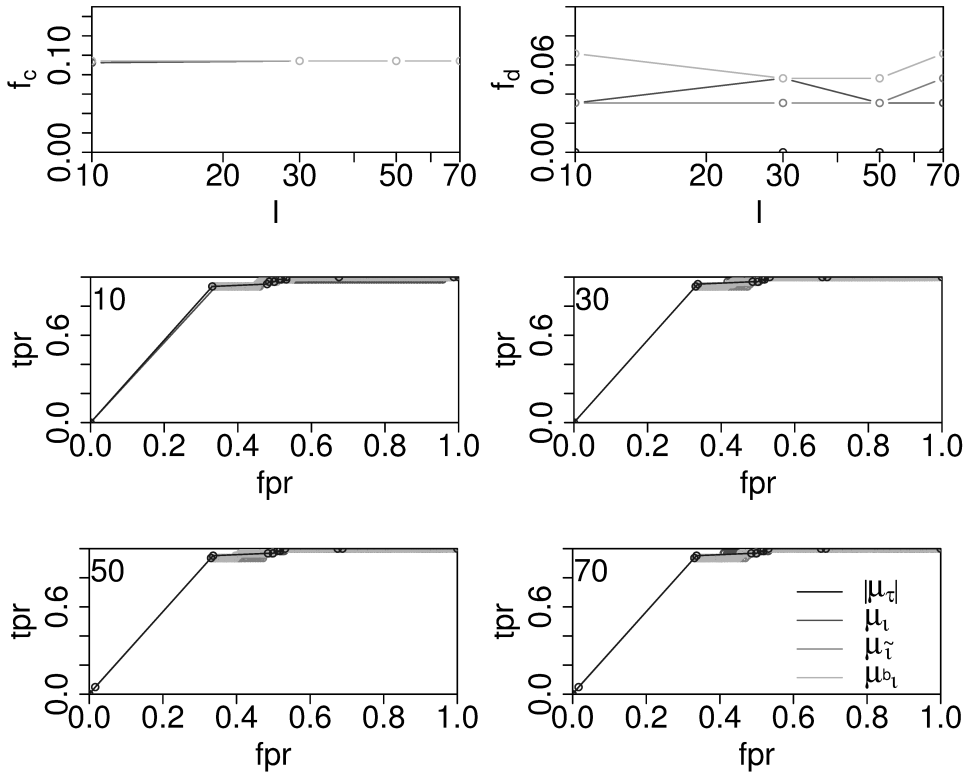

Figure 7.2: Dependence of $IOTA's$ reconstruction efficiency for a network of 60 genes in *E. coli* on the length of the time series (deterministic case). Fraction f_c of links where the coupling is identified correctly, and fraction f_d with correct coupling direction. For Kendall's τ f_d is zero per definition. Furthermore, ROC curves obtained for time series of different lengths (10, 30, 50, and 70 time points) are shown.

evaluated as a function of the length of the time series. In this case, $\mu(y^{(k)}, y^{(l)}) > \mu(y^{(l)}, y^{(k)})$ indicates a link directed from k to l in the actual network.

To examine the reconstruction efficiency, the resulting ROC curves (showing the tpr's and the fpr's while continuously tuning the threshold as explained in Section 2.2) are considered.

In the investigated (rather small) network and for short time series, as illustrated in Figs. 7.2 and 7.3 (upper left panels), the fraction f_c of true links associated with the largest values of the corresponding pairwise measure is generally small. This is due to the fact that several indirect links obtained also large values of the measure, which is reflected in the ratio between fpr and tpr (shown in the middle and lower panels as ROC curves obtained for time series of different lengths). Regarding the direction of coupling (upper right panels) it becomes apparent that the presence of small noise significantly improves the capability to correctly infer directionality, while in the noise-free case both directions are often indistinguishable.

It has to be noted, that in this example, the network reconstruction efficiency for all analyzed

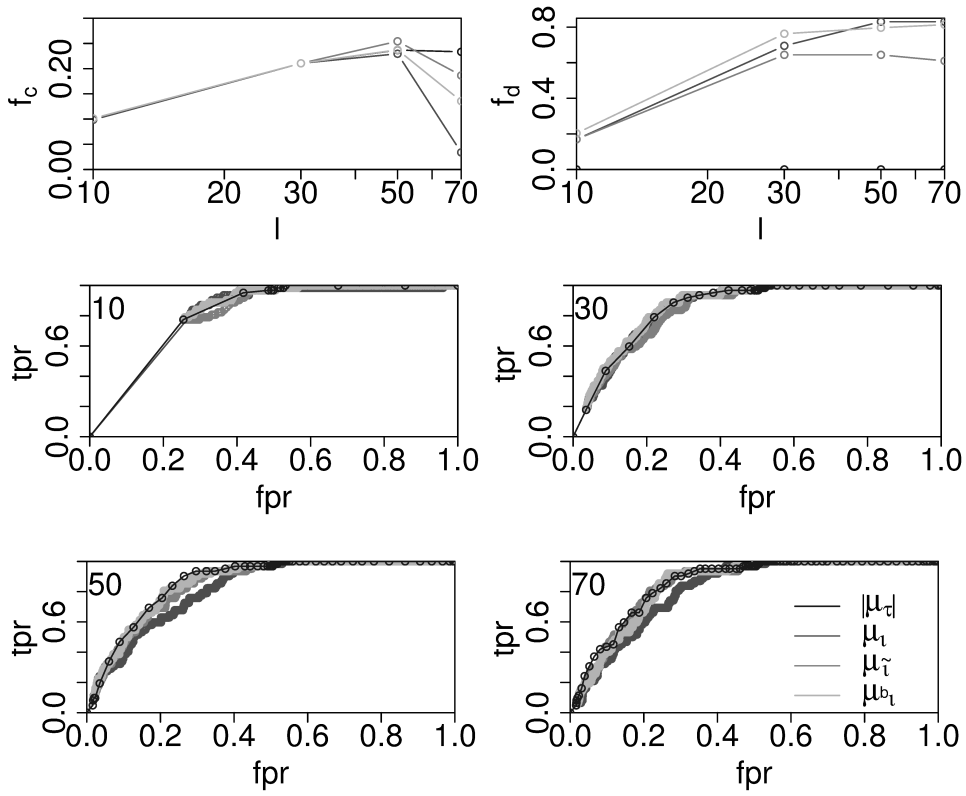

Figure 7.3: Results obtained for stochastic time series (noise level 0.01) of different lengths. Illustration analog to Fig. 7.2.

measures is limited rather by the complexity of the network than by the length of the time series. Hence, the obtained ROC curves have a very similar shape in all cases, in particular for the deterministic time series. However, Kendell's τ produces much more discontinuous curves than all versions of $IOTA$, which renders the network reconstruction with $IOTA$ more robust with respect to the choice of the threshold.

Furthermore, when considering a fixed threshold for the network reconstruction (*e.g.*, 0.99 as shown in Tab. 7.1 for the stochastic time series), a very similar efficiency of all $IOTA$ variants is obtained. This reconstruction strongly differs from the one obtained with Kendall's τ. In particular, for all considered lengths of the time series $IOTA$ predicts almost all links correctly, whereas approximately 50% of the false links can be excluded. On the other hand, the number of both, true and false identified links with Kendall's τ decreases with an increasing length of the time series. Thus, in contrast to $IOTA$ Kendall's τ predicts only an incomplete network, particularly from the longer time series.

Moreover, $IOTA$ determines the type of regulation (activation or inhibition) by using the sign of the average slope of the reordered time series $g^{(k,l)}$ (as explained in Chapter 5). Thus, a negative regulatory link within the considered network can be correctly distinguished from the

	no. of time points	10	30	50	70
ι	unidirectional links	61/62	60/62	60/62	60/62
	tpr	0.98	0.97	0.97	0.97
	fpr	0.51	0.50	0.50	0.50
$\tilde{\iota}$	unidirectional links	60/62	60/62	60/62	60/62
	tpr	0.97	0.97	0.97	0.97
	fpr	0.50	0.50	0.50	0.50
${}^b\iota$	unidirectional links	60/62	60/62	60/62	60/62
	tpr	0.97	0.97	0.97	0.97
	fpr	0.50	0.50	0.50	0.50
τ	tpr	0.77	0.60	0.56	0.44
	fpr	0.25	0.15	0.13	0.10

Table 7.1: Reconstruction efficiency of $IOTA$ for a gene regulatory subnetwork of 60 genes of *E. coli*, stochastic time series (noise level 0.01) of different length, and a threshold 0.99.

positive regulations, not only with Kendall's τ, but also with $IOTA$. Further tests with larger GRN's also revealed that although the type of interaction can not always be correctly inferred (some links were falsely identified to be inhibitory), the error rate is very similar to the one obtained when rank correlations, such as Kendall's τ, are applied.

7.1.2 Application to a network of 100 genes of E. coli

Next, $IOTA$ is applied to reconstruct the gene regulatory subnetwork of 100 genes of *E. coli*, which was also employed in the comparison study in Chapter 2 and 3. The investigated subnetwork is sparse, having 121 unidirectional links, 6 of which are autoregulatory. Moreover, the considered gene expression time series consist of 10 time points each, in order to compare the capabilities of $IOTA$ to infer networks particularly from very short gene expression time series with the performance of previously defined measures (discussed in Chapter 2). Both deterministic and stochastic time series are investigated. The study is carried out with μ_ι including the partial version of the measures and the significance test at significance level 0.01.

Weighting

First, the influence of different weighting functions (introduced in Tab. 5.1) is elucidated and the performance of $IOTA$ is compared to Kendall's τ on the basis of the resulting ROC curves.

Figure 7.4 illustrates that the range of the values of $IOTA$ depends strongly on the choice of the weighting function. The monotonicity of the reordered time series can be perturbed by external influences, presence of noise, or both, which may lead to fluctuations of the reordered time series. An uniform weighting renders the measure very sensitive to these influences since

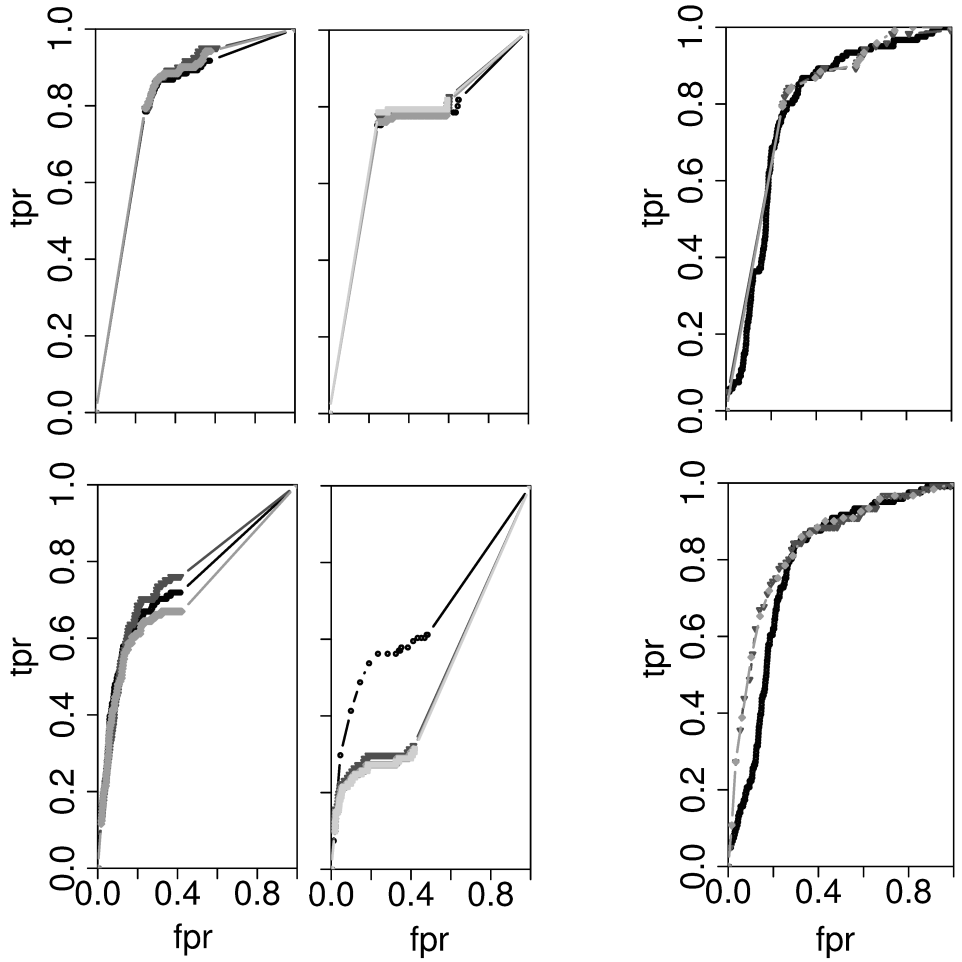

Figure 7.4: ROC curves for a network of 100 genes of *E. coli* (upper panels: noise-free case, lower panels: noise level of 0.1). The left and middle panels are obtained for $IOTA$ (superfluous links removed, permutation test included) for different weights (introduced in Tab. 5.1) – left: slope (black), squared slope (dark gray), maximal excursion (gray); middle: uniform (black), arithmetic mean (dark gray), geometric mean (gray), harmonic mean (light gray). The right panels belong to the correlations: Pearson (black), Spearman (dark gray), Kendall (gray). These results do not incorporate a significance test.

fluctuations lead to additional crossing points which highly increases the value of the sum in Eq. (5.2). However, the noise-induced fluctuations are expected to be small compared to those of the reordered time series of independent subsystems and, hence, they must be weighted less. It became apparent that all mean-based weights are less robust against the influence of noise than the slope-based ones. In particular, $IOTA$ has the lowest noise sensitivity using the squared slope weight – an important feature, especially when dealing with biological data.

By comparing the reconstruction efficiency of $IOTA$ to those of the rank correlations, the lower boundary of the fpr is similar in both cases, which poses a direct control on the false positives. However, the ROC curves for $IOTA$ are more continuous than those for the rank correlations. Hence, the network topology as inferred with $IOTA$ is less sensitive to the threshold chosen to decide which nodes to be linked. This is of particular use when dealing with experimental data.

Reconstruction scenarios

Next, I compare various reconstruction scenarios (using threshold 0.95 or 0.5, as well as deterministic and stochastic time series simulated at noise level 0.0 or 0.1, respectively) obtained with $IOTA$ (μ_ι with squared slope weighting function) and Kendall's τ. The results confirm that the network topology as inferred with $IOTA$ is less sensitive with respect to the threshold chosen for the network reconstruction.

While for a noise intensity of 0.0 a proper choice of the threshold is evident, it becomes problematic when stochasticity is involved, since the influence of the noise is difficult to quantify. For instance (Fig. 7.5), for noise intensity 0.1 and threshold 0.95, Kendall's rank correlation gives a tpr of less than 10% (fpr $\approx$ 1%), whereas a threshold of 0.5 renders a tpr $\approx$ 70% (fpr $\approx$ 25%). On the other hand, when $IOTA$ is applied under the same conditions, the tpr at both threshold levels is approximately 65% (fpr $\approx$ 30%(40%) for threshold 0.95(0.5)). Even though the best reconstruction efficiency under noisy conditions can be essentially obtained with Kendall's τ, the chance to achieve that optimum is small. In contrast to Kendall's τ, where the number of correctly and falsely identified links is strongly dependent on the threshold, the values obtained with $IOTA$ are almost constant. Thus, $IOTA$ results in robust predictions with respect to varying thresholds, demanded in practical applications. Furthermore, in contrast to Kendall's τ, which assumes all genes to be autoregulated per definition, $IOTA$ infers correctly all of the included autoregulatory links, and identifies partially genes which are not autoregulated (1% in the noise-free case and even 5% from the noisy time series).

Hidden time points

In the following, I briefly elucidate the influence of incomplete time series on $IOTA's$ reconstruction efficiency. Typically time-resolved measurements of gene expression are not only short, in addition, the data involves also missing values at distinct time points. Hence, a part of the data is hidden for the analysis and the full length of the time series is not available for all genes. Next, the subnetwork consisting of 100 genes in *E. coli* is revisited, where the nodes are described via simulated gene expression time series of 10 time points each. Noise is not considered here. However, from the 10×100 data points 1% is chosen at random to be hidden for the coupling analysis. Moreover, $IOTA$ is calculated only for time series with at least 8 matching time points, since otherwise the statistical significance would decrease to much compared to the time series with 10 matching time points. Again μ_ι is employed together with the partial variant of $IOTA$ and the permutation test (significance level 0.01) is performed.

The results shown in Fig. 7.6 indicate that hidding data points at random affects the reconstruction efficiency only little. While for the full lengths of all time series a true positives rate $tpr = 0.79$ is obtained, for the data set with hidden time points this value is reduced to

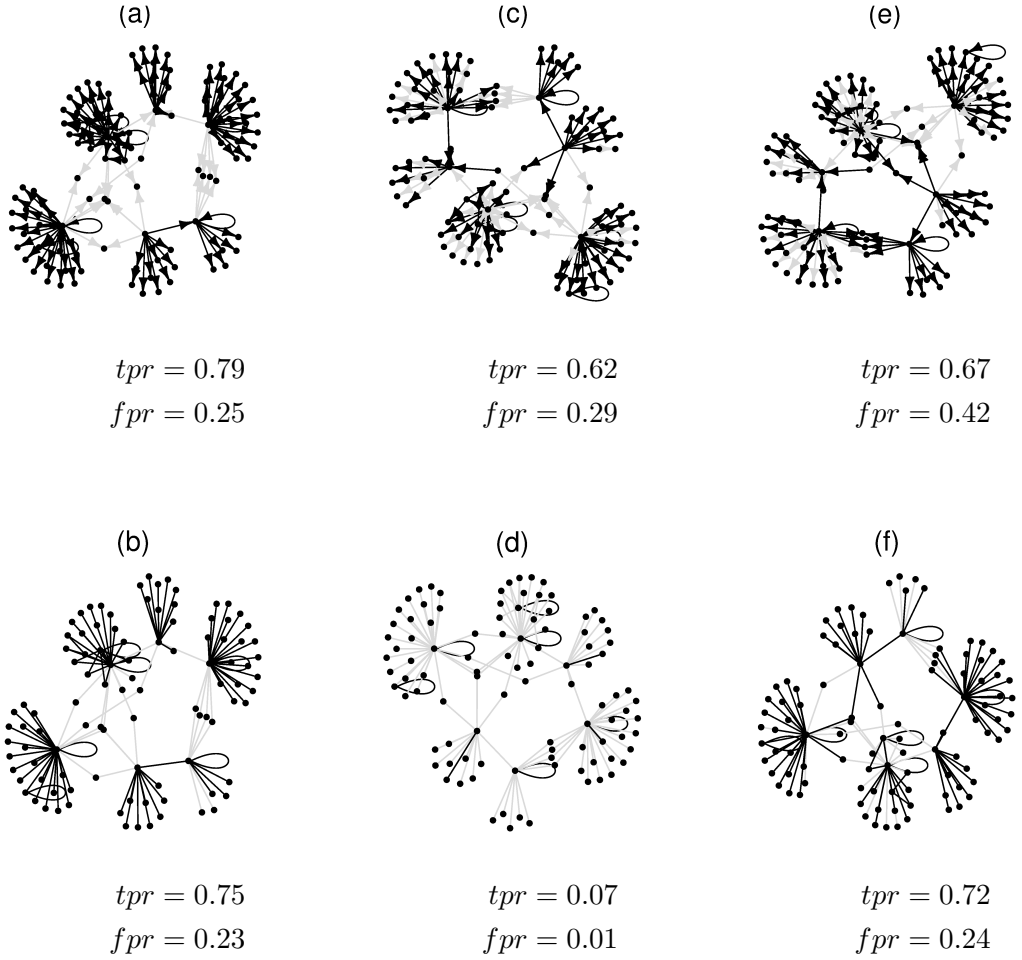

Figure 7.5: Reconstruction of a regulatory network of 100 genes of *E. coli* from (a)–(b) noise-free time series using threshold 1, (c)–(d) time series simulated with noise level 0.1 using threshold 0.95, and (e)–(f) with noise level 0.1 using threshold 0.5. (a),(c) and (e) show the networks obtained with $IOTA$, whereas (b),(d) and (f) are obtained with Kendall's τ. The original network (in the lower panels the undirected version) is shown in light gray, correctly identified links are marked in black. False positive links are not shown.

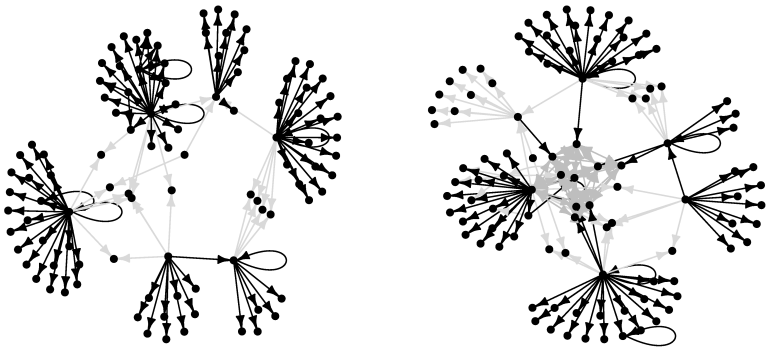

Figure 7.6: Reconstruction of the regulatory network from noise-free time series with $IOTA$ (threshold 1). The original network is shown in light gray, correctly identified links are marked in black. False positive links are not shown. Left panel: For each of the 100 genes the complete time series of length 10 time points is available for the analysis (1000 expression values). A true positives rate $tpr = 0.79$ and a false positives rate $fpr = 0.25$ are obtained in this case. Right panel: From the 1000 measured expression values 1% (randomly chosen) is not available for the analysis. $IOTA$ is calculated only for time series with at least 8 matching time points (pairs which do not fullfill this condition are marked in dark gray). This results in a true positives rate $tpr = 0.65$ and a false positives rate $fpr = 0.22$.

$tpr = 0.65$. However, this decrease mainly reflects the fact that $IOTA$ is not computed for all pairs of genes, since the number of matching time points is not always sufficient. The fpr's are even less affected.

7.1.3 Influence of the number of genes

Next, the influence of the size, and respectively the density of the network under study, on the reconstruction efficiency is examined in detail.

E. coli: In order to investigate the full capabilities of $IOTA$, the original source network of *E. coli* is modified to include additional bidirectional links (several of the existing unidirectional are replaced with bidirectional links). These links can have the same sign, either activating (ac–ac) or inhibitory (re–re) in both directions, corresponding to a positive feedback loop, or they can describe a negative feedback loop by having opposed signs (ac–re). Subnetworks of various sizes (Tab. 7.2) and time series of 10 time points each are generated with $SynTRen$, and μ_ι (together with the partial variant and including the permutation test at significance level 0.01) is applied to infer the networks.

The overall reconstruction efficiency of $IOTA$ is again displayed in terms of ROC curves. In Fig. 7.7 a lower boundary of the fpr can be observed, which decreases for increasing number of nodes, *i.e.*, decreasing network density. This boundary poses a direct control on the false positives. However, the observed decrease is not monotone which indicates that the network

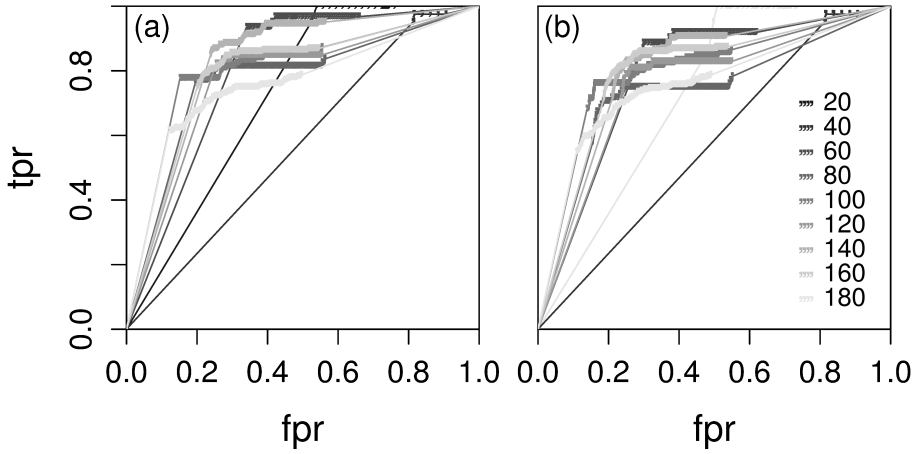

Figure 7.7: ROC curves for *IOTA* for networks of different sizes and (a) deterministic and (b) stochastic time series of 10 time points. In the case of stochastic time series a noise level of 0.01 is considered.

density is not alone governing the reconstruction efficiency. Other factors of influence might be the clustering, the occurrence of cascades, or the fraction of uni- and bidirectional coupling.

Additionally, it must be noted that for low noise level (0.01) the reconstruction efficiency is very similar to the one in the noise-free case. In particular, the shapes of the obtained ROC curves in the case of low noise intensity and in the noise-free case do not differ much from each other. Hence, the reconstructed network can be expected to be almost identical in both cases.

Moreover, in all cases the obtained ROC curves are rather continuous which supports once again that the network topology as inferred with *IOTA* is little sensitive to the threshold chosen to decide which nodes to be linked. This fact is additionally illustrated in Fig. 7.8 for various reconstruction scenarios in the low noise case (noise level 0.01). The reconstructed networks are obtained with *IOTA* when two different thresholds are used to identify the links, namely 0.75 and 0.99.

Furthermore, the reconstruction efficiency at the threshold of 0.99 is summarized in Tab. 7.2. Here, the first two small networks are too dense to get sufficiently low fpr's, while for larger networks the fpr can be reduced to approximately $20 - 30\%$. Furthermore, for sparse networks tpr's of approximately $70 - 80\%$ are usually achievable. However, the values of fpr and tpr are not monotonically increasing with decreasing density. They are additionally affected for instance by the local link density around the hubs. Moreover, the presence of bidirectional coupling can also have an effect on the reconstruction efficiency. Applied to networks of intermediate size (100 to 160 nodes), *IOTA* inferred approximately 50% of the bidirectional links present (Tab. 7.2 bidirectional). However, when applied to very short time series, *IOTA* tends to identify only one significant direction. Particularly for coupling strengths of opposed sign (as already shown in Section 6.1.2 for the small network modules), a bidirectional coupling can be obscured. Hence, increasing levels of bidirectional coupling can reduce the reconstruction efficiency.

Moreover, other association measures (*e.g.*, Kendall's τ) assume all genes to be autoregulated

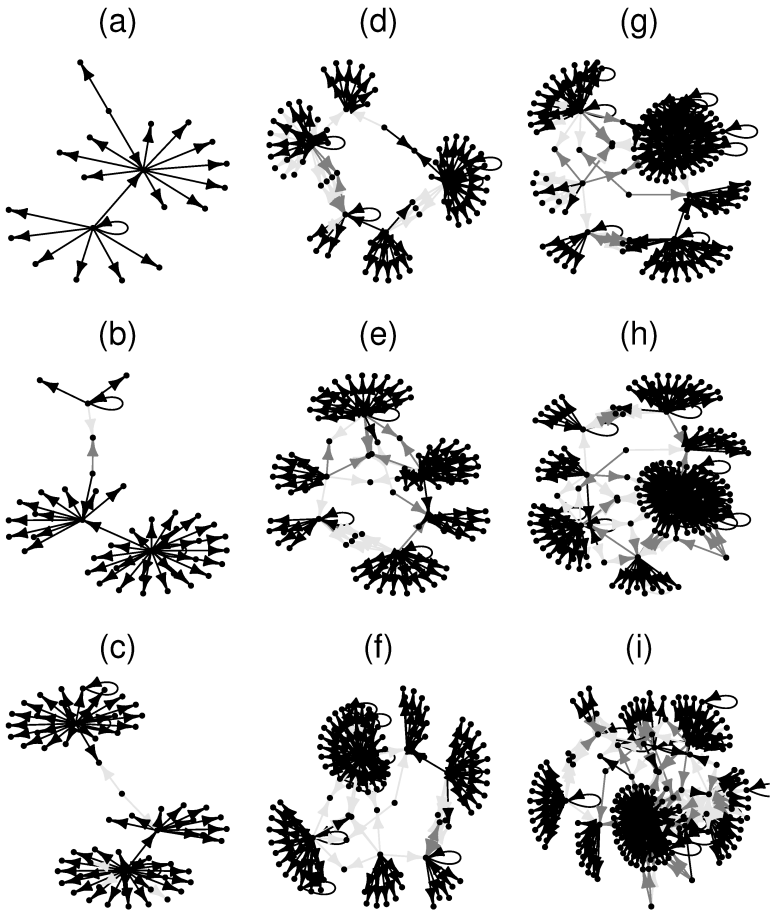

Figure 7.8: Network reconstruction with *IOTA* for networks of different sizes and stochastic time series (noise intensity 0.01) of 10 time points. Black links are obtained with threshold 0.75 (dark gray additionally with threshold 0.99); light gray indicates not identified links.

per definition. In contrast, *IOTA* infers correctly nearly all of the included autoregulatory links, but also identifies partially genes which are not autoregulated. This slightly reduces the fpr for several of the investigated networks (Tab. 7.2 autoregulated).

		(a)	(b)	(c)	(d)	(e)
no. of nodes		20	40	60	80	100
auto-regulation	yes	1/1	2/2	3/3	4/4	6/6
	no	0/19	0/38	0/57	1/76	0/94
unidirectional		20/20	39/41	60/62	73/89	98/121
bidirectional	*ac–ac*	0/0	0/0	0/0	0/0	1/1
	re–re	0/0	0/0	0/0	0/1	1/1
	re-ac	0/0	0/0	0/0	0/2	0/3
μ_ι ($\lvert\mu_\tau\rvert$)	*tpr*	1.00 (0.80)	0.95 (0.90)	0.92 (0.73)	0.73 (0.52)	0.79 (0.63)
	fpr	0.73 (0.43)	0.81 (0.79)	0.38 (0.24)	0.23 (0.14)	0.27 (0.13)

		(f)	(g)	(h)	(i)
no. of nodes		120	140	160	180
auto-regulation	yes	10/10	14/14	15/16	19/19
	no	0/110	0/126	14/144	0/161
unidirectional		124/147	160/179	170/210	179/255
bidirectional	*ac–ac*	0/1	1/1	1/3	0/3
	re–re	1/1	1/1	1/2	0/3
	re-ac	1/3	1/3	3/7	0/7
μ_ι ($\lvert\mu_\tau\rvert$)	*tpr*	0.81 (0.63)	0.84 (0.64)	0.79 (0.65)	0.65 (0.48)
	fpr	0.28 (0.16)	0.27 (0.20)	0.21 (0.17)	0.19 (0.10)

Table 7.2: Reconstruction efficiency of $IOTA$, threshold 0.99, for *E. coli*.

S. cerevisiae: As a second example the regulatory network of *S. cerevisiae* is considered. The source network is supplemented with bidirectional links and subnetworks of various sizes (Tab. 7.3) as well as time series of 10 time points each are generated with *SynTRen*.

Both, *E. coli* and *S. cerevisiae*, have similar network properties (*e.g.*, both show approximately a power-law behavior of the out-degree distribution. The mean degree is 2.9 for the *E. coli* and 3.3 for the *S. cerevisiae* network, and the clustering coefficient is 0.024 for the *E. coli* and 0.016 for the *S. cerevisiae* network). However, the systems differ strongly in the dynamics of the simulated time series (Fig. 7.9). In particular, the gene expression has more constant values in the case of the *S. cerevisiae* network, which impedes the network reconstruction problem as illustrated in Fig. 7.10. This is true for both, $IOTA$ and Kendall's τ. Nontheless, the general tendencies which were observed for the *E. coli* subnetworks are also reproduced with the regulatory network of *S. cerevisiae* as shown in (Tab. 7.3).

7.2 Reconstructing a GRN of *Chlamydomonas reinhardtii* from experimental data

Eventually, I employ the relevance network approach combined with $IOTA$ to reconstruct a GRN for *Chlamydomonas reinhardtii* using experimentally obtained data sets.

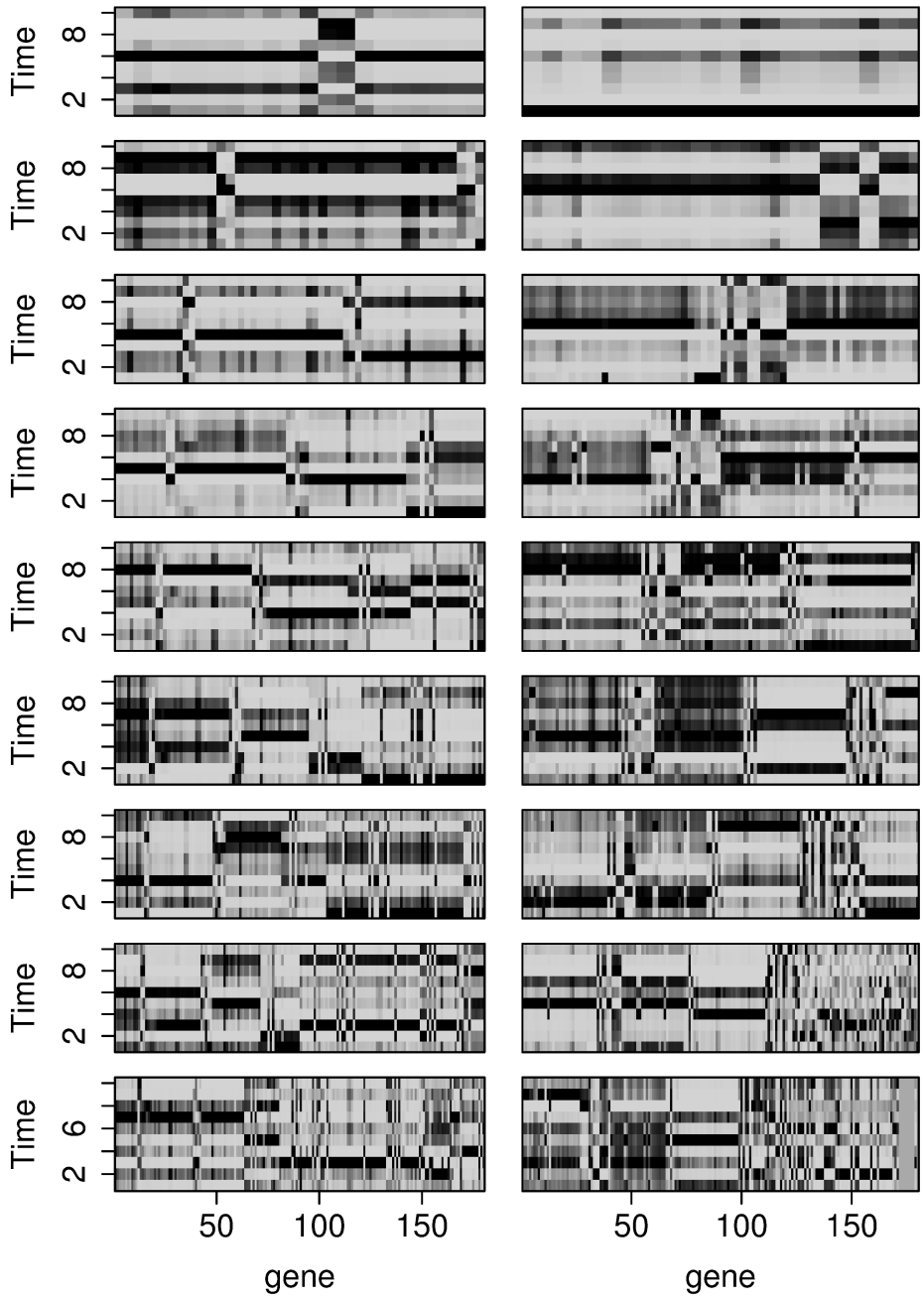

Figure 7.9: Illustration of the normalized expression rates simulated for *E. coli* (left panels) and *S. cerevisiae* (right panels) regulatory networks of different sizes with noise level of 0.01. The network sizes correspond to those in Tab. 7.2 and Tab. 7.3 with 20 nodes in the upper and 180 in the lower panel.

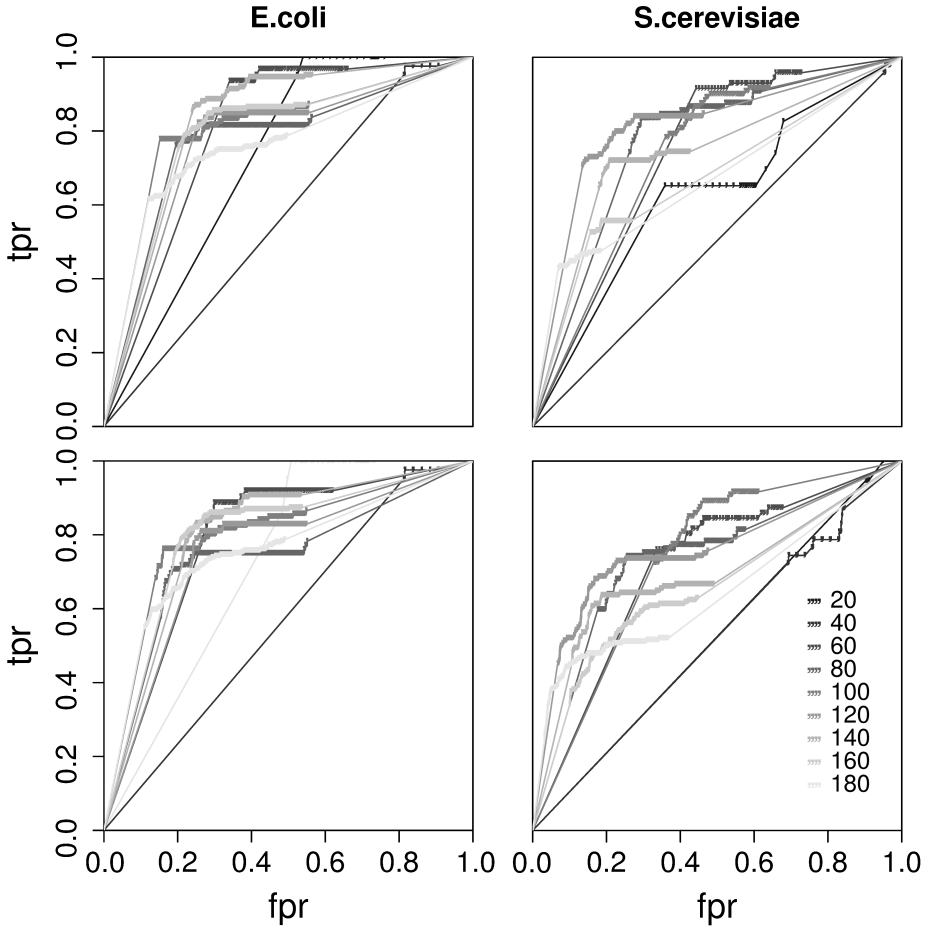

Figure 7.10: ROC curves for *IOTA* for networks of different sizes and deterministic (upper panels) and stochastic (lower panels) time series of 10 time points. In the case of stochastic time series a noise level of 0.01 is considered.

		(a)	(b)	(c)	(d)	(e)
no. of nodes		20	40	60	80	100
auto-regulation	yes	0/0	0/0	0/0	0/0	0/0
	no	20/20	0/40	0/60	0/80	0/100
unidirectional		0/18	33/35	54/59	64/85	86/108
bidirectional	*ac–ac*	0/0	1/1	0/1	0/1	0/1
	re–re	0/0	0/3	0/3	2/3	1/3
	re-ac	0/0	1/2	2/2	1/2	2/2
μ_ι ($\lvert\mu_\tau\rvert$)	*tpr*	1.00 (0.95)	0.87 (0.72)	0.87 (0.48)	0.75 (0.34)	0.78 (0.51)
	fpr	0.95 (0.90)	0.84 (0.58)	0.66 (0.29)	0.25 (0.15)	0.39 (0.27)

		(f)	(g)	(h)	(i)
no. of nodes		120	140	160	180
auto-regulation	yes	0/0	0/0	0/0	0/0
	no	45/120	75/140	15/160	52/180
unidirectional		91/128	96/150	83/172	93/195
bidirectional	*ac–ac*	1/1	0/1	0/1	0/2
	re–re	0/3	0/3	0/3	0/3
	re-ac	0/4	0/5	1/7	0/7
μ_ι ($\lvert\mu_\tau\rvert$)	*tpr*	0.68 (0.34)	0.61 (0.35)	0.46 (0.21)	0.46 (0.19)
	fpr	0.17 (0.10)	0.17 (0.10)	0.16 (0.10)	0.12 (0.03)

Table 7.3: Reconstruction efficiency of $IOTA$, threshold 0.99, for *S. cerevisiae*.

Chlamydomonas reinhardtii (short *C. reinhardtii*) is a photosynthetic, unicellular, eukaryotic green alga frequently used as a model organism for cell and molecular biology, since it can be cultivated under controlled conditions and unicellularity precludes any influence of tissue heterogeneity or developmental factors. Moreover, it is widely distributed in soil and fresh water all over the world, the full genome sequence is known, plenty of mutants exist and genetic manipulation is relatively easy compared to other organisms. Cultivated under controlled conditions in a bioreactor *C. reinhardtii* is employed to uncover various biological processes (*e.g.*, related to photosynthesis and carbon metabolism), and to analyze the regulation in response to selected environmental factors. In that context, the study of the carbon concentrating mechanism (CCM) is of particular interest here.

Preceding analyses of *C. reinhardtii's* CCM have already revealed the identity of several genes which govern important regulatory functions. However, a detailed analysis of the expression patterns of transcription factors involved in the regulation of the CCM was lacking for a long time.

Thus, in order to identify the key regulators, $IOTA$ is applied to gene expression data from the following experimental setup [VWAMRP$^+$]: Cells of *C. reinhardtii* were cultured under photoautotrophic and temperature controlled conditions and with a continuous supply of light (photosynthetic photon flux density $PPFD \approx 200 \frac{\mu E}{m^2 s}$) and CO_2 (5% CO_2 in air). After the cell culture reached an optical density of 0.5 at 750 nm (approximately $3 \cdot 10^6$ cells per mL) it

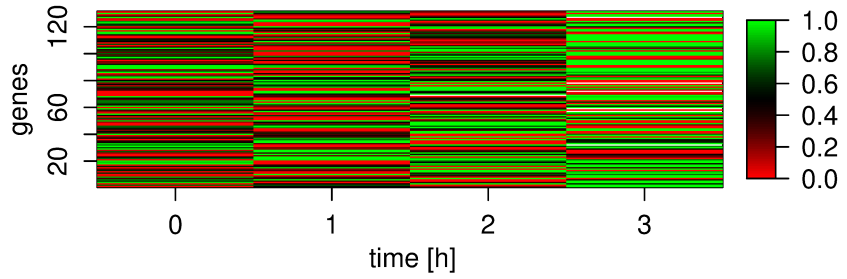

Figure 7.11: Observed expression pattern of transcription factor and transcription regulator coding genes involved in the CCM during carbon depreciation.

was sampled and the concentration of CO_2 was reduced to 0.04%. Further samples of the cell culture were collected 1, 2 and 3 hours after CO_2 reduction.

The medians of five biological replicates were available for the analysis, leading to gene expression time series which consist of 4 time points each and are uniformly sampled in time. The expression of 131 transcription factor and transcription regulator coding genes was monitored (Fig. 7.11), 7 of which result in incomplete time series and were excluded from the analysis.

The pairwise and partial $IOTA$ measures are applied to the gene expression data and the statistical analysis is based on the previously described permutation test, where the empirical p-values are estimated at significance level 0.01. Moreover, in order to reconstruct the regulatory network for the CCM a threshold of 0.95 is chosen.

This approach allows to reverse-engineer the topology of the CCM regulatory network, which exhibits a complex structure as shown in Fig. 7.12 (left panel), including previously known low-CO_2-responsive regulates (such as the transcription factor LCR1) [VWAMRP$^+$]. The inferred network has a rather different distribution of the incoming and the outgoing links (Fig. 7.13), however, the average degree is 4.15, both for the in-degree and the out-degree. That means, each transcription factor or transcription regulator in the inferred GRN regulates on average four genes and is regulated by four genes on average.

The previous study for evaluating the performance of $IOTA$ on synthetic data sets has demonstrated that approximately 75% of the uni- and bidirectional links were correctly identified, while the fpr's were fixed around 25%. Thus, this analyis represents a first step towards the understanding of the CCM. Having the restrictions in mind, the reconstructed network, can serve as a basis for designing specific experiments, since it indicates new target genes and reveals candidates that might govern the regulation of the CCM.

For instance, the inferred GRN predicts that $Lcr1$ (the gene that codes for the already known low-CO_2-responsive transcription factor LCR1) regulates five other transcription factor or transcription regulator coding genes. Those genes are members of the following transcription factor or transcription regulator families: C2C2-Dof, SBP, Orphan, FHA and C3H. Additionally, two genes are identified to potentially regulate a common subgroup of those genes predicted to be regulated by $Lcr1$, namely genes coding for (i) a MYB-related transcription factor, and (ii) a transcription regulator from the SNF2 family. The corresponding regulatory subnetwork is

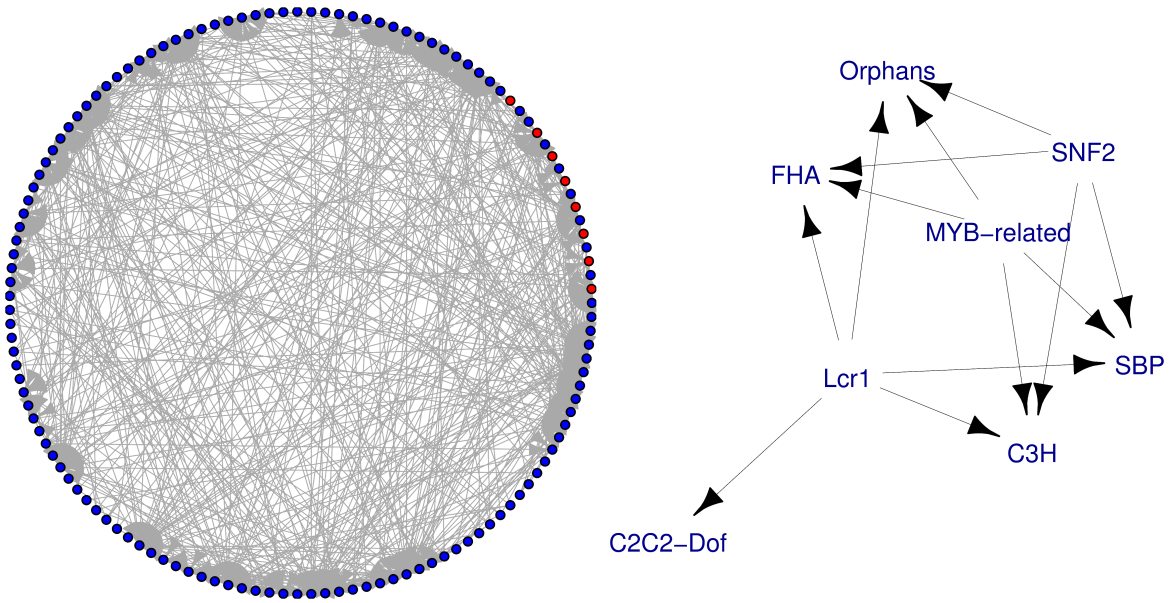

Figure 7.12: Left: The GRN of the CCM inferred from gene expression pattern during carbon depreciation exhibits a complex structure. Right: Gene regulatory subnetwork of the CCM during carbon depreciation corresponding to the nodes which are marked in red in the left panel.

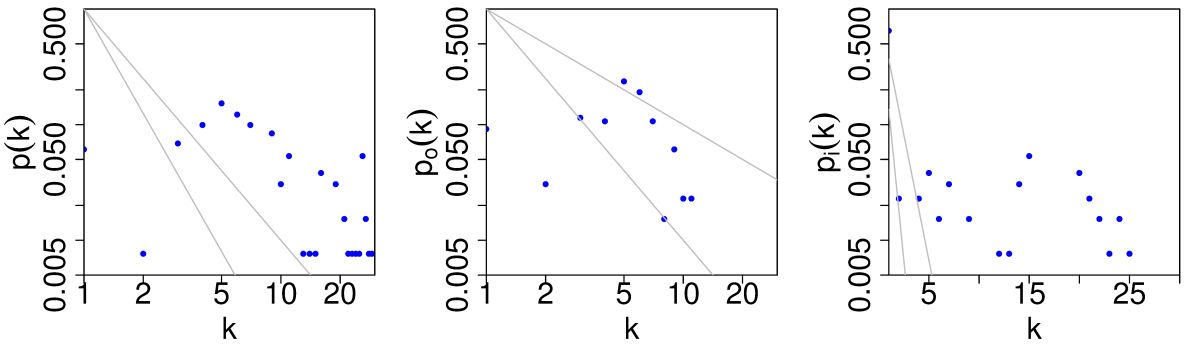

Figure 7.13: Degree distribution of the GRN of the CCM inferred from gene expression pattern during carbon depreciation (left: total degree, middle: out-degree, right: in-degree). The gray lines are to guide the eye towards the distribution which is expected from literature.

shown in Fig. 7.12 (right panel).

The MYB-related transcription factor coding gene, from here onwards referred to as *Lcr*2, is predicted to regulate four of the five genes that putatively are regulated by *Lcr*1, namely SBP, Orphan, FHA and C3H. Furthermore, a motif search analysis revealed similar sequence motifs and hence possible cis-regulatory elements in the promoter regions of *Lcr*1 and *Lcr*2, suggesting that both genes may be co-regulated by similar upstream *DNA*-binding proteins [VWAMRP$^+$].

Additionally, the SNF2 transcription regulator coding gene has been identified as a potential regulator of the same subgroup of genes that are regulated by *Lcr*1 and *Lcr*2, where the expression of these putative target genes is repressed after 3 hours.

It must be noted that further experimental and theoretical studies are needed to gain full insight into the regulatory mechanisms of CCM. Nevertheless, the analysis based on *IOTA* has already revealed several regulatory links, which would be worth exploring in further detailed investigations.

8 Conclusion

Understanding the functionality of an organism and its adaptation mechanisms to changing environmental conditions is one of the crucial topics of complex systems analysis. This requires comprehensive mechanistic, qualitative and quantitative insights into the different GRN's. However, in contrast to typical situations in physics, the measurements of gene expression are usually very limited, with short (typically 4 to 12 time points) and noisy time series. In addition, the sampling in time is often coarse and not uniform, and only few realizations of the experiment are available. On the other hand, the number of interacting elements (genes) ranges, in general, from hundreds to thousands.

Nevertheless, time series analysis can be the first step for the inference of GRN's. To this end, various association measures and network reconstruction tools are applied to time series measurements in order to provide first insights into the regulatory structure. Although, these network reconstructions frequently suffer from large false positive rates, they can serve as a basis to design more specific experiments.

In this context, I examined the relevance network approach as a flexible tool for reverse engineering GRN's from short time-resolved data (*i.e.*, approximately 10 time points). By performing a comprehensive comparison study on the basic relevance network approach using 21 different measures, I exposed that with a suitable choice of the association measure, this reverse engineering method is applicable to short time series. However, most of the currently used association measures have highly limited capabilities, as the number of time points that is usually available from gene expression measurements is in many cases not sufficient to infer the underlying structure of the network. This in turn makes the distinction between direct and indirect interactions an even more challenging task.

My results indicated that **rank and symbol based measures** have a significantly better performance in inferring interdependencies, whereas most of the standard measures (such as Granger causality and several information-theoretic measures) fail when very short time series are considered. Thus, the standard association measures (based directly on the time series) are often not suitable for GRN reconstruction. It is necessary to move towards measures rooted in the study of symbolic dynamics or ranks deduced from the time series, in order to increase the efficiency of the relevance network algorithm.

Although measures based on symbolic dynamics performed significantly well in the noise-free case, their performance was decreased as the noise level in the system increased, and for high noise intensities the reconstruction efficiency became comparable to that of mutual information. This implies that in the presence of strong noise, rank correlations are the most efficient tools for GRN reconstruction, since their performance was not significantly affected as the noise level increased.

Furthermore, using 6 scoring schemes together with different association measures, I showed that the extended relevance network algorithm can be used to improve the network reconstruc-

tion. In particular, I introduced two **novel asymmetric scoring schemes**, since most of the association measures, including the rank and symbol based measures, are symmetric, *i.e.*, the directionality of the interactions cannot be inferred. On the other hand, the performance of the few asymmetric measures (*i.e.*, Granger causality) was deficient for the short time series under consideration. I showed that a **novel scoring scheme**, denoted as the **asymmetric weighting (*AWE*)**, stands as a valuable approach to overcome the problems of introducing directionality in the reconstruction of the regulatory networks.

This study can serve as a basis for the selection of a reverse engineering method for network reconstruction, based on the combination of an association measure and a scoring scheme suitable for given data.

Moreover, I introduced **inner composition alignment (*IOTA*)**, a **novel** permutation-based measure and variants thereof, as efficient tools to identify relations between subsystems, together with the associated directionality, without the need of additional scoring. The measure has the following merits:

- *IOTA* is applicable to infer statistically significant (nonlinear) couplings from very short time series.
- It is capable to infer bi- or unidirectional coupling together with its directionality (in particular if the dynamics of the coupled subsystems involve small time delays).
- The approach allows to infer the type of regulation (activation or inhibition).
- The partial measure can distinguish indirect from direct coupling and indicate autoregulation.

Thus, *IOTA* is the only existing association measure which can determine **all** necessary characteristics when reconstructing regulatory networks.

In an extensive numerical study, I investigated the performance of *IOTA* to infer couplings within various networks which represent different dynamical systems. In particular, I showed that this new measure outperforms the correlation measures when it is applied to short time series. In this context, *IOTA* was applied to gene expression measurements including synthetic and experimental data; however, the reconstruction efficiency was also tested for autoregressive processes and chaotic oscillators in phase-coherent and non-phase-coherent regime, with promissing results.

Moreover, I showed that the reconstruction of GRN's with *IOTA* is not very sensitive to the choice of the threshold, which is used to define a link. This is of particular value when dealing with noisy experimental data, since noise disturbs the similarity of the time series, which renders the proper choice of a threshold problematic for several measures, such as rank correlations.

Finally, I used the relevance network approach with *IOTA* as the association measure in order to infer the GRN of green algae of the species *Chlamydomonas reinhardtii* under carbon deprivation. The reconstructed network was used to indicate candidate genes which play a role in the regulation of the carbon concentration mechanism. This in turn can serve as a basis to design more specific experiments. Those analyses will yield an improved understanding of the carbon concentration mechanism in *Chlamydomonas reinhardtii* and in plants in general.

In future, a dynamical modeling of the mechanism shall follow in order to predict the adaption of plants to the changing environmental conditions. This, however, requires additional experiments to further reduce the number of false positives, more detailed analysis of the network topology and longer time series measurements to investigate the dynamical parameters.

Futhermore, since $IOTA$ has been shown to be a valuable tool for coupling analysis, the measure shall be further applied to other (short) experimentally obtained gene expression data (*e.g.*, to infer the GRN that governs the early development of multicellular organisms, such as sea urchin or sea anemone). In addition, the application of the measure to time-resolved data from other coupled dynamical systems, such as electro-chemical oscillators is planned, which also requires a more detailed analysis of the performance of $IOTA$ when reconstructing relations between different coupled chaotic oscillatators (*e.g.*, Roesler, Lorenz, Van der Pol, or Duffing oscillators). In this context, in particular, the influence of transient behavior requires detailed investigation.

Bibliography

[AAN04] Albert, R.; Albert, I.; Nakarado, G. L.: Structural vulnerability of the north american power grid. In: *Physical Review E*, volume 69:p. 025103, 2004.

[AB02] Albert, R.; Barabasi, A. L.: Statistical mechanics of complex networks. In: *Reviews of Modern Physics*, volume 74:pp. 47–97, 2002.

[AC01] Aach, J.; Church, G. M.: Aligning gene expression time series with time warping algorithms. In: *Bioinformatics*, volume 17(6):pp. 495–508, 2001.

[Aka03] Akaike, H: A new look at the statistical model identification. In: *IEEE Transactions on Automatic Control*, volume 19(6):pp. 716–723, 2003.

[Alb05] Albert, R.: Scale-free networks in cell biology. In: *Journal of Cell Science*, volume 118(21):pp. 4947–4957, 2005.

[Alo07] Alon, U.: Network motifs: theory and experimental approaches. In: *Nature Reviews Genetics*, volume 8(6):pp. 450–461, 2007.

[AO03] Albert, R.; Othmer, H. G.: The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in drosophila melanogaster. In: *Journal of Theoretical Biology*, volume 223(1):pp. 1–18, 2003.

[AVB07] Almaas, E.; Vazquez, A.; Barabasi, A. L.: *Scale-free networks in biology*, chapter 1, pp. 1–19. World Scientific Publishing Company, 1st edition, 2007.

[BJ04] Bar-Joseph, Z.: Analyzing time series gene expression data. In: *Bioinformatics*, volume 20:pp. 2493–2503, 2004.

[BK00] Butte, A. J.; Kohane, I. S.: Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. In: *Pacific Symposium on Biocomputing On-Line Proceedings*, volume 5, pp. 415–426. 2000.

[BLM+06] Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D.-U.: Complex networks: Structure and dynamics. In: *Physics Reports*, volume 424(4-5):pp. 175–308, 2006.

[BRV01] Barabasi, A. L.; Ravasz, E.; Vicsek, T.: Deterministic scale-free networks. In: *Physica A: Statistical Mechanics and its Applications*, volume 299(3-4):pp. 559–564, 2001.

[BS89] Bogacki, P.; Shampine, L.F.: A 3(2) pair of runge-kutta formulas. In: *Applied Mathematics Letters*, volume 2(4):pp. 321–325, 1989.

[BS03] Brazma, A.; Schlitt, T.: Reverse engineering of gene regulatory networks: a finite state linear model. In: *Genome Biology*, volume 4(6):p. P5, 2003.

[BS05] Balazs, P.; Stephen, O.: Genome-wide analysis of context-dependence of regulatory networks. In: *Genome Biology*, volume 6(2):p. 206, 2005.

[BSS04] Baba, K.; Shibata, R.; Sibuya, M.: Partial correlation and conditional correlation as measures of conditional independence. In: *Australian and New Zealand Journal of Statistics*, volume 46(4):pp. 657–664, 2004.

[BTS+00] Butte, A. J.; Tamayo, P.; Slonim, D.; Golub, T. R.; Kohane, I. S.: Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. In: *Proceedings of the National Academy of Sciences*, volume 97:pp. 12182–12186, 2000.

[CC07] Capitani, P.; Ciaccia, P.: Warping the time on data streams. In: *Data and Knowledge Engineering*, volume 62(3):pp. 438–458, 2007.

[CLM04] Crucitti, P.; Latora, V.; Marchiori, M.: A topological analysis of the italian electric power grid. In: *Physica A: Statistical Mechanics and its Applications*, volume 338(1-2):pp. 92 – 97, 2004.

[CPB+02] Caiani, E.; Porta, A.; Baselli, G.; Turiel, M.; Muzzupappa, S.; Pagani, M.; Malliani, A.; Cerutti, S.: Analysis of cardiac left-ventricular volume based on time warping averaging. In: *Medical and Biological Engineering and Computing*, volume 40(2):pp. 225–233, 2002.

[DAB03] Doiron, B.; A., Longtin; B., Lindner: Oscillatory network coding of a global stimulus. volume 201. 2003.

[DCB06] Ding, Mingzhou; Chen, Yonghong; Bressler, Steven: Granger causality: Basic theory and application to neuroscience, 2006. URL `http://arxiv.org/abs/q-bio/0608035v1`.

[DCB07] Ding, M.; Chen, Y.; Bressler, S. L.: Granger causality: Basic theory and application to neuroscience. In: Schelter, B.; Winterhalder, M.; Timmer, J., editors, *Handbook of Time Series Analysis*, pp. 437–460. Wiley InterScience, 2007.

[dJ02] de Jong, H.: Modeling and simulation of genetic regulatory systems: A literature review. In: *Journal of Computational Biology*, volume 9(1):pp. 67–103, 2002.

[DNR77] Dempster, A.P.; N.M., Laird.; Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. In: *Journal of the Royal Statistical Society, Series B*, volume 39(1):pp. 1–38, 1977.

[DRO+02] Davidson, E. H.; Rast, J. P.; Oliveri, P.; Ransick, A.; Calestani, C.; Yuh, C. H.; Minokawa, T.; Amore, G.; Hinman, V.; Arenas-Mena, C.; Otim, O.; Brown, C. T.; Livi, C. B.; Lee, P. Y.; Revilla, R.; Rust, A. G.; Pan, Z.; Schilstra, M. J.; Clarke, P. J.; Arnone, M. I.; Rowen, L.; Cameron, R. A.; McClay, D. R.; Hood, L.; Bolouri, H.: A genomic regulatory network for development. In: *Science*, volume 295(5560):pp. 1669–1678, 2002. ISSN 1095-9203.

[DWFS98] D'haeseleer, P.; Wen, X.; Fuhrman, S.; Somogyi, R.: Mining the gene expression matrix: inferring gene relationships from large scale gene expression data. In: *Proceedings of the second international workshop on Information processing in cell and tissues*, pp. 203–212. Plenum Press, Sheffield, United Kingdom, 1998.

[DZD+10] Donner, R. V; Zou, Y. Z.; Donges, J. F.; Marwan, N.; Kurths, J.: Recurrence networks – a novel paradigm for nonlinear time series analysis. In: *New Journal of Physics*, volume 12(3):p. 033025, 2010.

[DZMK09] Donges, J. F.; Zou, Y.; Marwan, N.; Kurths, J.: The backbone of the climate network. In: *EPL*, volume 87(4):p. 48007, 2009.

[ESBB98] Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D.: Cluster analysis and display of genome-wide expression patterns. In: *Proceedings of the National Academy of Sciences*, volume 95(25):pp. 14863–14868, 1998.

[Faw06] Fawcett, T.: An intoduction to ROC analysis. In: *Pattern Recognition Letters*, volume 27(8):pp. 861–874, 2006.

[FC06] Ferre, F.; Clote, P.: BTW: a web server for Boltzmann Time Warping of gene expression time series. In: *Nucleic Acids Res*, volume 34:pp. 482–485, 2006.

[FFF+03] Förster, J; Famili, I.; Fu, P.; Palsson, B..; Nielsen, J.: Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. In: *Genome Research*, volume 13(2):pp. 244–253, 2003.

[FFR05] Fluss, R.; Faraggi, D.; Reiser, B.: Estimation of the Youden index and its associated cutoff point. In: *Biometrical Journal*, volume 47(4):pp. 458–472, 2005.

[FP07] Frenzel, S.; Pompe, B.: Partial mutual information for coupling analysis of multivariate time series. In: *Physical Review Letters*, volume 99(20):pp. 204101–, 2007.

[GBBK02] Guelzim, N.; Bottani, S.; Bourgine, P.; Kepes, F.: Topological and causal structure of the yeast transcriptional regulatory network. In: *Nature Genetics*, volume 31(1):pp. 60–63, 2002.

[Gio09] Giorgino, T.: Computing and visualizing dynamic time warping alignments in R: The dtw package. In: *Journal of Statistical Software*, volume 31(7):pp. 1–24, 2009.

[Gio10] Giorgino, T.: *Dynamic Time Warping algorithms*. CRAN, Feb. 2010. Package version 1.14-3.

[GR02] Gibbons, F. D.; Roth, F. P.: Judging the quality of gene expression-based clustering methods using gene annotation. In: *Genome Research*, volume 12(10):pp. 1574–1581, 2002.

[GSK+08] Guo, S.; Seth, A. K.; Kendrick, K. M.; Zhou, C.; Feng, J.: Partial Granger causality – eliminating exogenous inputs and latent variables. In: *Journal of Neuroscience Methods*, volume 172(1):pp. 79–93, 2008.

[HKKN11] Hempel, S.; Koseska, A.; Kurths, J.; Nikoloski, Z.: Inner composition alignment for inferring directed networks from short time series. In: *Physics Review Letters*, volume 107:p. 054101, 2011.

[HKNa] Hempel, S.; Koseska, A.; Nikoloski, Z.: Data-driven reconstruction of directed networks. In: *Physical Review E*. Submitted, 15 Aug 2011.

[HKN+b] Hempel, S.; Koseska, A.; Nikoloski, Z.; Kiss, I.; Kurths, J.: Inferring direct and indirect interactions in networks of chaotic oscillators. In preparation.

[HKNK11] Hempel, S.; Koseska, A.; Nikoloski, Z.; Kurths, J.: Unraveling gene regulatory networks from time-resolved gene expression data – a measures comparison study. In: *BMC Bioinformatics*, volume 12:p. 292, 2011.

[HKT99] Hegger, R.; Kantz, H.; T., Schreiber: Practical implementation of nonlinear time series methods: The TISEAN package. In: *CHAOS*, volume 9(2):pp. 413–435, 1999.

[HKY99] Heyer, L. J.; Kruglyak, S.; Yooseph, S.: Exploring expression data: Identification and analysis of coexpressed genes. In: *Genome Research*, volume 9(11):pp. 1106–1115, 1999.

[HSPVB07] Hlavackova-Schindler, K.; Palus, M.; Vejmelka, M.; Bhattacharya, J.: Causality detection based on information-theoretic approaches in time series analysis. In: *Physics Reports*, volume 441(1):pp. 1–46, 2007.

[IG09] Ihaka, R.; Gentleman, R.: R version 2.9.2, 2009.

[Joh67] Johnson, S. C.: Hierarchical clustering schemes. In: *Psychometrika*, volume 32(3):pp. 241–254, 1967.

[JW02] Jackson, M. O.; Watts, A.: The evolution of social and economic networks. In: *Journal of Economic Theory*, volume 106(2):pp. 265–295, 2002.

[Kau69] Kauffman, S. A.: Metabolic stability and epigenesis in randomly constructed genetic nets. In: *Journal of Theoretical Biology*, volume 22:pp. 437– 467, 1969.

[Koh82] Kohonen, T.: Self-organized formation of topologically correct feature maps. In: *Biological Cybernetics*, volume 43(1):pp. 59–69, 1982.

[LBY+04] Luscombe, N. M.; Babu, M. M.; Yu, H.; Snyder, M.; Teichmann, S. A.; Gerstein, M.: Genomic analysis of regulatory network dynamics reveals large topological changes. In: *Nature*, volume 431(7006):pp. 308–312, 2004. ISSN 1476-4687.

[LFS98] Liang, S.; Fuhrman, S.; Somogyi, R.: Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In: *Pacific Symposium on Biocomputing On-Line Proceedings*, volume 3, pp. 18 – 29. 1998.

[Li90] Li, W.: Mutual information functions versus correlation functions. In: *Journal of Statistical Physics*, volume 60(5):pp. 823–837, 1990.

[LRR+02] Lee, T. I.; Rinaldi, N. J.; Robert, F.; Odom, D. T.; Bar-Joseph, Z.; Gerber, G. K.; Hannett, N. M.; Harbison, C. T.; Thompson, C. M.; Simon, I.; Zeitlinger, J.; Jennings, E. G.; Murray, H. L.; Gordon, D. B.; Ren, B.; Wyrick, J. J.; Tagne, J. B.; Volkert, T. L.; Fraenkel, E.; Gifford, D. K.; Young, R. A.: Transcriptional regulatory networks in Saccharomyces cerevisiae. In: *Science*, volume 298(5594):pp. 799–804, October 2002. ISSN 1095-9203.

[LT03] Levine, M.; Tjian, R.: Transcription regulation and animal diversity. In: *Nature*, volume 424(6945):pp. 147–151, 2003.

[LW08] Liang, K. C.; Wang, X.: Gene regulatory network reconstruction using conditional mutual information. In: *EURASIP Journal on Bioinformatics and Systems Biology*, volume 2008:p. 14, 2008.

[Mac] MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: *Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press.

[MC07] Mukhopadhyay, N. D.; Chatterjee, S.: Causality and pathway search in microarray time series experiment. In: *Bioinformatics*, volume 23(4):pp. 442–449, 2007.

[Mey09] Meyer, P. E.: *infotheo: Information-Theoretic Measures*. CRAN, 2009. R package version 1.1.0.

[MKLB07] Meyer, P. E.; Kontos, K.; Lafitte, F.; Bontempi, G.: Information-theoretic inference of large transcriptional regulatory networks. In: *EURASIP Journal on Bioinformatics and Systems Biology*, volume 2007:p. 9, 2007.

[MLB09] Meyer, P. E.; Lafitte, F.; Bontempi, G.: *minet: Mutual Information Network Inference*. CRAN, 2009. R package version 2.0.0.

[MRTK07] Marwan, N.; Romano, M. C.; Thiel, M.; Kurths, J.: Recurrence plots for the analysis of complex systems. In: *Physics Reports*, volume 438(5-6):pp. 237–329, 2007.

[MSOI+02] Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; Chklovskii, D.; Alon, U.: Network motifs: Simple building blocks of complex networks. In: *Science*, volume 298(5594):pp. 824–827, 2002.

[NCARR10] Nepomuceno-Chamorro, I.; Aguilar-Ruiz, J.; Riquelme, J.: Inferring gene regression networks with model trees. In: *BMC Bioinformatics*, volume 11(1):p. 517, 2010.

[NDSS11] Neusius, T.; Daidone, I.; Sokolov, I. M.; Smith, J. C.: Configurational subdiffusion of peptides: A network study. In: *Physical Review E*, volume 83:p. 021902, 2011.

[NF08] Noe, F.; Fischer, S.: Transition networks for modeling the kinetics of conformational changes in macromolecules. In: *Current Opinion in Structural Biology*, volume 18(2):pp. 154 – 162, 2008.

[NRT+10] Nawrath, J.; Romano, M. C.; Thiel, M.; Kiss, I. Z.; Wickramasinghe, M.; Timmer, J.; Kurths, J.; Schelter, B.: Distinguishing direct from indirect interactions in oscillatory networks with multiple time scales. In: *Physical Review Letters*, volume 104(3):pp. 038701–, 2010.

[OMWL08] Osterhage, H.; Mormann, F.; Wagner, T.; Lehnertz, K.: Detecting directional coupling in the human epileptic brain: Limitations and potential pitfalls. In: *Physical Review E*, volume 77:p. 011914, 2008.

[PBL+11] Parlitz, U.; Berg, S.; Luther, S.; Schirdewan, A.; Kurths, J.; Wessel, N.: Classifying cardiac biosignals using ordinal pattern statistics and symbolic dynamics. In: *Computional Biology and Medicine*, volume 19:p. 21511252, 2011.

[Pea85] Pearl, J.: Bayesian networks: A model of self-activated memory for evidential reasoning. In: *Proceedings of the 7th Conference of the Cognitive Science Society*, pp. 329–334. University of California Press, 1985.

[Pfa08a] Pfaff, B.: *Analysis of Integrated and Cointegrated Time Series with R*. Springer, New York, 2nd edition, 2008. ISBN 0-387-27960-1.

[Pfa08b] Pfaff, B.: Var, svar and svec models: Implementation within R package vars. In: *Journal of Statistical Software*, volume 27(4):pp. 1–32, 2008.

[PKHS01] Palus, M.; Komarek, V.; Hrncir, Z.; Sterbova, K.: Synchronization as adjustment of information rates: Detection from bivariate time series. In: *Physical Review E*, volume 63(4):pp. 046211–, 2001.

[PRK01] Pikovsky, A.; Rosenblum, M.; Kurths, J.: *Synchronization: a universal concept in nonlinear sciences*. Cambridge University Press, Cambridge, UK, 1st edition, 2001. ISBN 0-521-53352-X.

[QQKKG02] Quian Quiroga, R.; Kraskov, A.; Kreuz, T.; Grassberger, P.: Performance of different synchronization measures in real data: A case study on electroencephalographic signals. In: *Physical Review E*, volume 65:p. 041903, 2002.

[RP01] Rosenblum, M. G.; Pikovsky, A. S.: Detecting direction of coupling in interacting oscillators. In: *Physical Review E*, volume 64:p. 045202, Sep 2001.

[RPK96] Rosenblum, M. G.; Pikovsky, A. S.; Kurths, J.: Phase synchronization of chaotic oscillators. In: *Physical Review Letter*, volume 76:pp. 1804–1807, 1996.

[RRSA02] Ronen, M.; Rosenberg, R.; Shraiman, B. I.; Alon, U.: Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. In: *Proceedings of the National Academy of Sciences*, volume 99(16):pp. 10555–10560, 2002. ISSN 0027-8424.

[SBA07] Soranzo, N.; Bianconi, G.; Altafini, C.: Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. In: *Bioinformatics*, volume 23(13):pp. 1640–1647, 2007.

[SC78] Sakoe, H.; Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. In: *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume ASSP-26(1):pp. 43–49, 1978.

[Sch00] Schreiber, T.: Measuring information transfer. In: *Physical Review Letters*, volume 85(2):pp. 461–, 2000.

[SMC07] Stolovitzky, G.; Monroe, D.; Califano, A.: Dialogue on reverse-engineering assessment and methods: The dream of high-throughput pathway inference. In: *Annals of the New York Academy of Sciences*, volume 1115(1):pp. 1–22, 2007.

[SMF11] Schaffter, T.; Marbach, D.; Floreano, D.: Genenetweaver: In silico benchmark generation and performance profiling of network inference methods. In: *Bioinformatics*, volume 27(16):pp. 2263–2270, 2011.

[SOMMA02a] Shen-Orr, S. S.; Milo, R.; Mangan, S.; Alon, U.: Network motifs in the transcriptional regulation network of Escherichia coli. In: *Nature Genetics*, volume 31(1):pp. 64–68, 2002.

[SOMMA02b] Shen-Orr, S. S.; Milo, R.; Mangan, S.; Alon, U.: Network motifs in the transcriptional regulation network of Escherichia coli. In: *Nature Genetics*, volume 31(1):pp. 64–68, 2002.

[SSP09] Stewart, A. J.; Seymour, R. M.; Pomiankowski, A.: Degree dependence in rates of transcription factor evolution explains the unusual structure of transcription networks. In: *Proceedings of the Royal Society B: Biological Sciences*, volume 276(1666):pp. 2493–2501, 2009.

[SW92] Smith, D. A.; White, D. R.: Structure and dynamics of the global economy: Network analysis of international trade 1965–1980. In: *Social Forces*, volume 70(4):pp. 857–893, 1992.

[SWD+06] Schelter, B.; Winterhalder, M.; Dahlhaus, R.; Kurths, J.; Timmer, J.: Partial phase synchronization for multivariate synchronizing systems. In: *Physical Review Letters*, volume 96:p. 208103, 2006.

[TGQS09] Tormene, P.; Giorgino, T.; Quaglini, S.; Stefanelli, M.: Matching incomplete time series with dynamic time warping: An algorithm and an application to post-stroke rehabilitation. In: *Artificial Intelligence in Medicine*, volume 45(1):pp. 11–34, 2009.

[VdBVLN+06a] Van den Bulcke, T.; Van Leemput, K.; Naudts, B.; van Remortel, P.; Ma, H.; Verschoren, A.; De Moor, B.; Marchal, K.: SynTReN: A generator of synthetic gene expression data for design and analysis of structure learning algorithms. In: *BMC Bioinformatics*, volume 7(1):p. 43, 2006.

[VdBVLN+06b] Van den Bulcke, T.; Van Leemput, K.; Naudts, B.; van Remortel, P.; Ma, H.; Verschoren, A.; De Moor, B.; Marchal, K.: SynTRen generator, 03 2006. Version 1.1.3.

[Voh01] Vohradsky, J.: Neural network model of gene expression. In: *FASEB Journal*, volume 15(3):pp. 846–854, 2001.

[VVNV08] Veiga, D.; Vicente, F.; Nicolas, M.; Vasconcelos, A. T.: Predicting transcriptional regulatory interactions with artificial neural networks applied to e. coli multidrug resistance efflux pumps. In: *BMC Microbiology*, volume 8(1):p. 101, 2008.

[VWAMRP+] Vischi Winck, F.; Arvidsson, S.; Mauricio Riano-Pachon, D.; Hempel, S.; Koseska, A.; Nikoloski, Z.; Rupprecht, J.; Mueller-Roeber, B.: Deciphering the gene regulatory network of the green alga Chlamydomonas reinhardtii under carbon deprivation. In: *BMC Systems Biology*. Submitted, 18 Aug 2011.

[VZ70] Velichko, V. M.; Zagoruyko, N. G.: Automatic recognition of 200 words. In: *International Journal of Man-Machine Studies*, volume 2(3):pp. 223–234, 1970.

[WFM+98] Wen, X.; Fuhrman, S.; Michaels, G. S.; Carr, D. B.; Smith, S.; Barker, J. L.; Somogyi, R.: Large-scale temporal gene expression mapping of central nervous system development. In: *Proceedings of the National Academy of Sciences*, volume 95(1):pp. 334–339, 1998.

[WS98] Watts, D. J.; Strogatz, S. H.: Collective dynamics of "small-world" networks. In: *Nature*, volume 393:pp. 440–442, 1998.

[WSR+09] Wessel, N.; Suhrbier, A.; Riedl, M.; Marwan, N.; Malberg, H.; Bretthauer, G.; Penzel, T.; J., Kurths: Detection of time-delayed interactions in biosignals using symbolic coupling traces. In: *Europhysics Letters*, volume 87(1):pp. 10004–, 2009.

[WSS02] Warren, C. P.; Sander, L. M.; Sokolov, I. M.: Geography in a scale-free network model. In: *Physical Review E*, volume 66:p. 056105, 2002.

[WZV+04] Wille, A.; Zimmermann, P.; Vranova, E.; Furholz, A.; Laule, O.; Bleuler, S.; Hennig, L.; Prelic, A.; von Rohr, P.; Thiele, L.; Zitzler, Ec.; Gruissem, W.; Buhlmann, P.: Sparse graphical gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. In: *Genome Biology*, volume 5(11):p. R92, 2004.

[XBS02] Xulvi-Brunet, R.; Sokolov, I. M.: Evolving networks with disadvantaged long-range connections. In: *Physics Review E*, volume 66:p. 026118, 2002.

[YP11] Yu, D.; Parlitz, U.: Inferring network connectivity by delayed feedback control. In: *PLOS ONE*, volume 6(9):p. e24333, 2011.

[ZLS+06] Zhang, Z.; Liu, C.; Skogerb, G.; Zhu, X.; Lu, H.; Chen, L.; Shi, B.; Zhang, Y.; Wang, J.; Wu, T.; Chen, R.: Dynamic changes in subgraph preference profiles of crucial transcription factors. In: *PLoS Computational Biology*, volume 2(5):pp. e47+, 2006.

[ZZXS10] Zhang, J.; Zhou, C.; Xu, X.; Small, M.: Mapping from structure to dynamics: A unified view of dynamical processes on networks. In: *Physical Review E*, volume 82(2):p. 026116, 2010.

[ZZZL+07] Zhou, C.; Zemanova, L.; Zamora-Lopez, G.; Hilgetag, C. C.; Kurths, J.: Structure-function relationship in complex brain networks expressed by hierarchical synchronization. In: *New Journal of Physics*, volume 9, 2007.

List of Figures

List of Tables

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Berlin, den 27. Januar 2012 Sabrina Hempel

Im Rahmen dieser Dissertation entstanden die folgenden Veröffentlichungen:

- Unraveling gene regulatory networks from time-resolved gene expression data – a measures comparison study [HKNK11]
- Inner composition alignment for inferring directed networks from short time series [HKKN11]
- Data-driven reconstruction of directed networks [HKNa]

Danksagung

An dieser Stelle möchte ich allen meinen Dank aussprechen, die zum Gelingen dieser Arbeit beigetragen haben.

Zunächst danke ich meinem Doktorvater Prof. Kurths für die Möglichkeit unter seiner Betreuung an der Humboldt Universität zu Berlin promomieren und in seiner Arbeitsgruppe am PIK in Potsdam dieses interessante Forschungsthema bearbeiten zu können. Ich danke für die Bereitstellung der Arbeitsmittel, die Freiräume, die Diskussionen und die hilfreichen Hinweise.

Mein Dank gilt aber auch der gesamten Arbeitsgruppe für den interessanten Austausch und das stets angenehme Arbeitsklima. Insbesondere danke ich Dr. Koseska für die intensive Zusammenarbeit und die nützlichen Hinweise und Anregungen im Zusammenhang mit dieser Arbeit. Auch danke ich meinem Bruder Dr. Reik Donner für zahlreiche nützliche Diskussionen.

Weiterhin möchte ich mich bei den Kollegen am MPI für Pflanzenphysiology und an der Universität Potsdam für die Zusammenarbeit bedanken. Mein besonderer Dank gilt hier Prof. Müller-Röber, Dr. Vischi Winck und Dr. Arvidson für die Bereitstellung der experimentellen Daten und die Zusammenarbeit bei deren Analyse. Darüber hinaus danke ich Dr. Nikoloski für die Zusammenarbeit und die ausführlichen Diskussionen zum theoretischen Teil meiner Arbeit.

Schließlich danke ich meiner ganzen Familie für die Unterstützung und Geduld. Mein besonderer Dank gilt meinem Mann Martin für seine Gelassenheit, die Diskussionen und nützlichen Hinweise.